

Exascale and then what...

A View of Post-Exascale Computational Science
and the Emerging Mix of HPC, AI and Quantum



Crescat scientia; vita excolatur

Rick Stevens
Argonne National Laboratory
The University of Chicago

Some thoughts – Exa/AI/Qu/Zetta \Rightarrow EAZQ

- **Exascale** – how did we get here and what do still have to deliver to make good on the promises we made
- **AI** – huge opportunity that is impacting nearly everything we will do going forward, but we need to own the AI and Science coupling
- **Quantum** – we probably need to curb the enthusiasm a bit, but what should we be doing to make history proud of us?
- **Zetta** – Progress towards Zetta needs to underpin the overall plan as the deep foundations need reinforcement and Q will not replace it

The image features a complex, futuristic background. It consists of a dense network of glowing, multi-colored lines (primarily blue, purple, and teal) that resemble a circuit board or a data network. These lines are set against a dark, starry space background with subtle nebulae and light flares. In the center of the image, a large, bold, white letter 'E' is prominently displayed, serving as the focal point. The overall aesthetic is high-tech and digital.

Summer of 2007

Modeling and Simulation at the Exascale for Energy and the Environment

Co-Chairs:
Horst Simon
*Lawrence Berkeley National Laboratory
April 17-18, 2007*
Thomas Zacharia
*Oak Ridge National Laboratory
May 17-18, 2007*
Rick Stevens
*Argonne National Laboratory
May 31-June 1, 2007*

Office of Science
U.S. DEPARTMENT OF ENERGY

Early of 2008

**ExaScale Computing Study:
Technology Challenges in
Achieving Exascale Systems**

Peter Kogge, Editor & Study Lead

**Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snively
Thomas Sterling
R. Stanley Williams
Katherine Yelick**

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724**. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.



Leadership Facility Strategy for the future Roadmap and Timeline

Through a three-phase plan executed over the next decade, the LCF will deploy a series of ever-more-powerful, balanced, scalable, HPC and data resources to support the most challenging computational problems of the nation.

Phase 1: Procure and operate pre-Exascale systems (2018) Three-way RFP w/ ORNL, ANL, LLNL

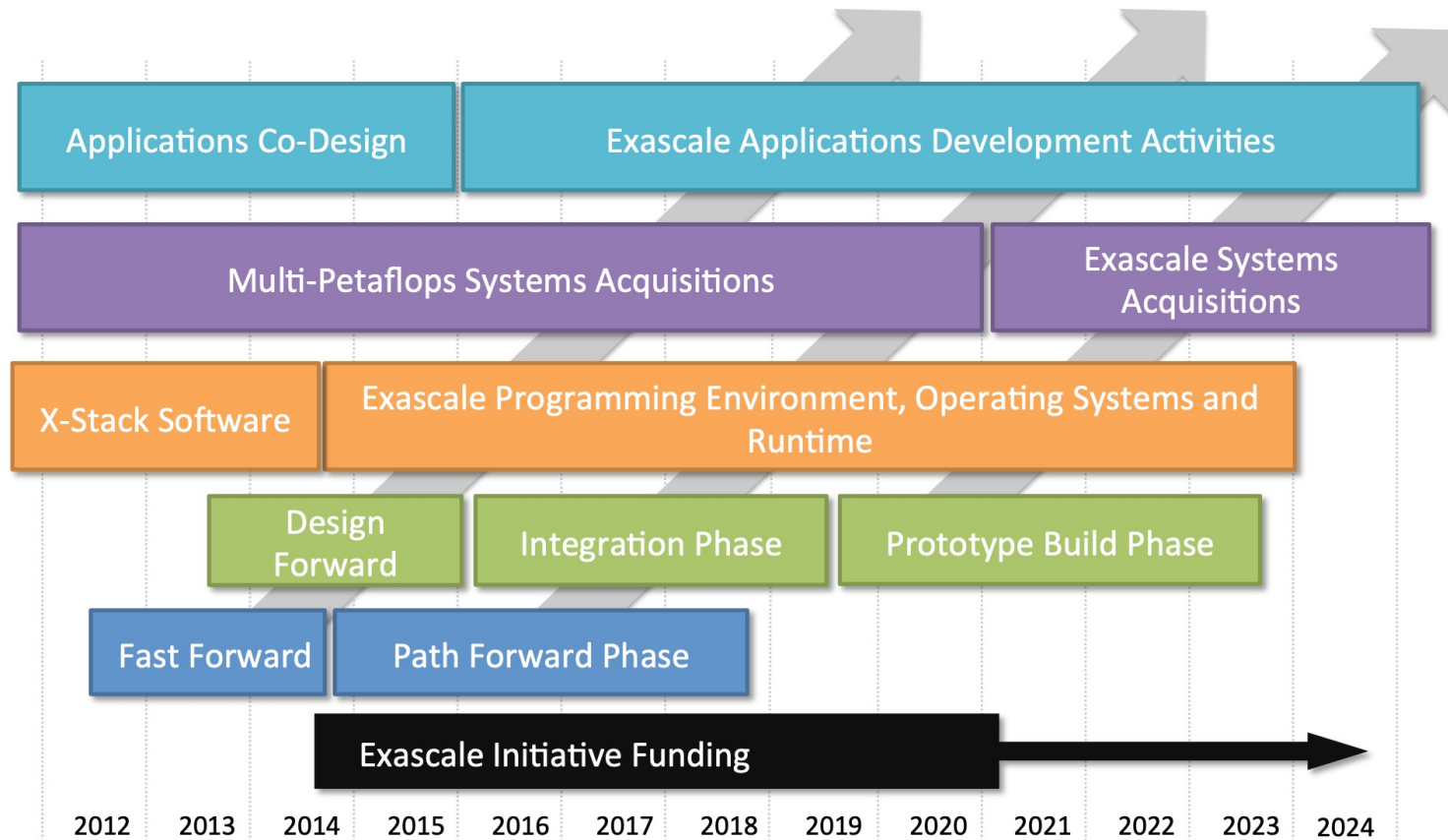
Phase 2: Procure and operate Exascale systems (2022)

Phase 3: Procure and operate Second generation Exascale systems (2026)

Computer System requirements for each Leadership Computing Center

	2012	2018	2022	2026
Peak FLOP/s	10-20 PF	100-200 PF	500-2000 PF	2000-4000 PF
Memory	0.5-1 PB	5-10 PB	32-64 PB	50-100 PB
I/O Buffer	N/A	500 TB	3 PB	5 PB
Storage Disk +tape	20+100 PB	100+1000 PB	1+10 EB	5+50 EB
Power & Space	6-12 MW 5,000-10,000 ft ²	15-20 MW 8,000-15,000 ft ²	20-30 MW 20,000 ft ²	25-35 MW 25,000 ft ²

Exascale Program Timeline



Pre-exascale and Exascale US Landscape

System	Delivery	CPU + Accelerator Vendor
Summit	2018	IBM + NVIDIA V100
Sierra	2018	IBM + NVIDIA V100
Perlmutter	2021	AMD + NVIDIA A100
Polaris	2021	AMD + NVIDIA A100
Frontier	2021	AMD + AMD MI250x
Crossroads	2023	Intel
Aurora	2023	Intel + Intel PVC
El Capitan	2023	AMD + AMD MI300



Aurora

Leadership Computing Facility
Exascale Supercomputer
Overview

Peak Performance
≥ 2 Exaflops DP

Intel GPU

**Intel® Data Center GPU Max
Series 1550**
Code named "PVC"

Intel Xeon PROCESSOR

**Intel® Xeon® CPU Max Series with
HBM**
Code named "SPR+HBM"

Platform

HPE Cray-Ex

System Size

166 Compute Racks
10,624 nodes
21,248 CPUs
63,744 GPUs

Compute Node

2 CPU, 6 GPU
1 TB DDR5
1 TB HBM
8 Fabric NICs
Node Unified Memory Architecture

Aggregate System Memory

DDR5 10.9 PB, 5.95 PB/s
HBM CPU 1.36 PB, 30.5 PB/s
HBM GPU 8.16 PB, 208.9 PB/s

System Interconnect

HPE Slingshot 11
Dragonfly topology with adaptive
routing
2.12 PB/s Peak Injection BW
0.69 PB/s Peak Bisection BW

High-Performance Storage

220 PB
31 TB/s DAOS bandwidth
1024 DAOS nodes

Programming Environment

oneAPI
C/C++
Fortran
SYCL/DPC++
Python
Aurora MPICH and oneCCL
OpenMP offload
Kokkos, RAJA
Intel PerformanceTools, Intel gdb
Tensorflow, PyTorch
DDP, Horovod, DeepSpeed
oneDAL and ScikitLearn
Python Libraries
JupyterHub
Julia, Numba
Spark
MLDE, SmartSim





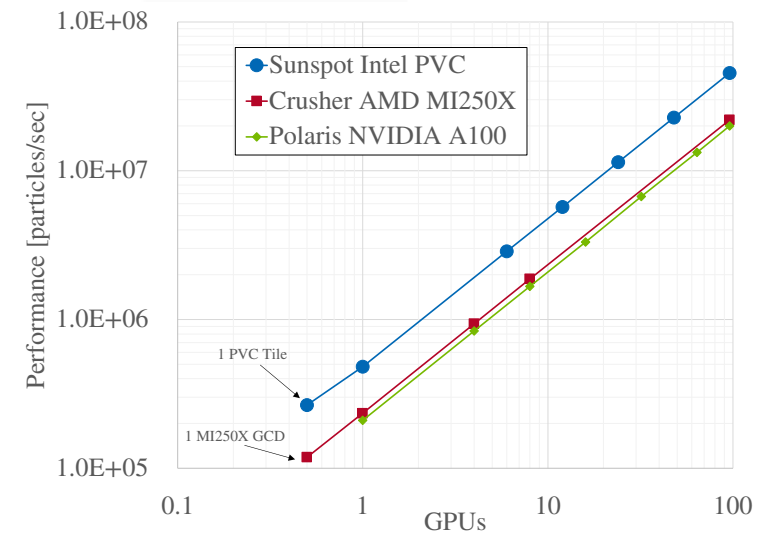
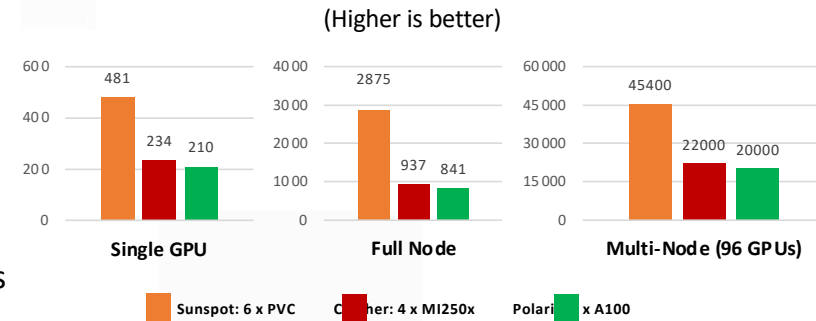
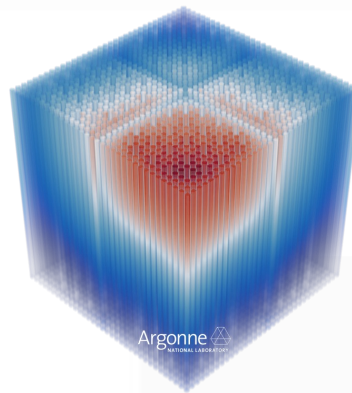
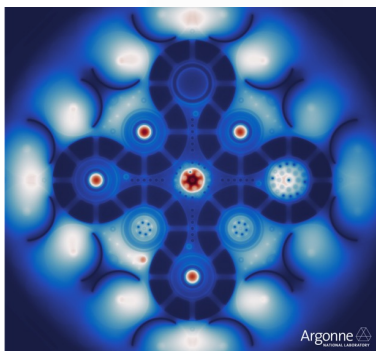
Delivering on Exascale Science

⇒ Large Number of Applications

OpenMC (courtesy of John Tramm)

<https://docs.openmc.org>

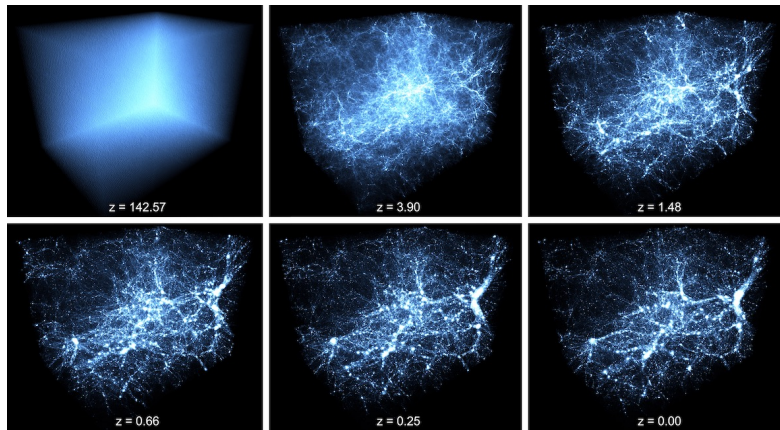
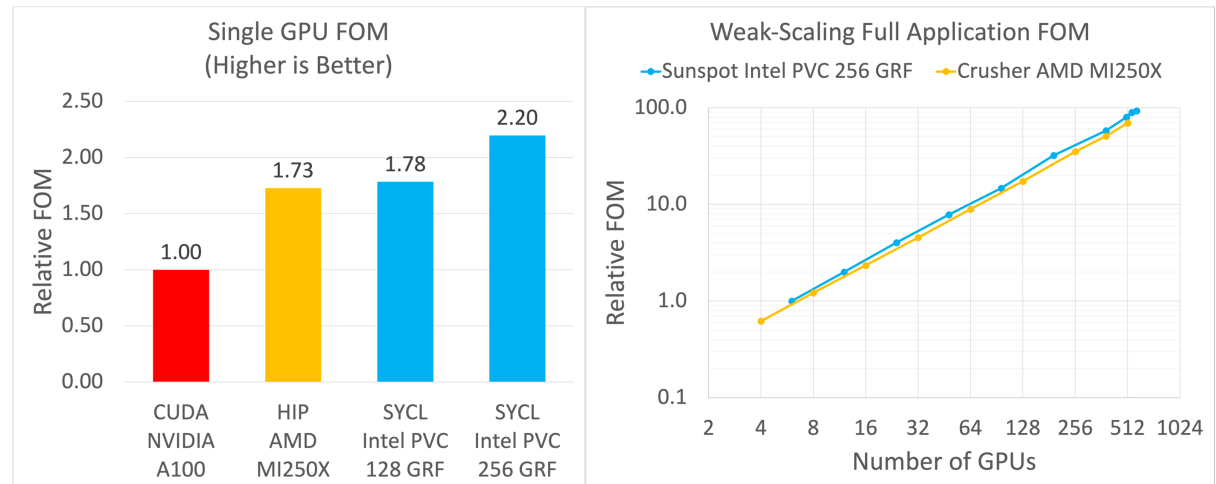
- OpenMC is being developed as part of the ECP ExaSMR project (PIs: Steven Hamilton, Paul Romano)
- OpenMC is a Monte Carlo particle transport code written in C++ and the OpenMP target offloading programming model
- The project seeks to accelerate the design of small modular nuclear reactors by generating virtual reactor simulation datasets with high-fidelity, coupled physics models for reactor phenomena that are truly predictive
- The Monte Carlo method employed by OpenMC is considered the "gold standard" for high-fidelity but these methods suffer from a very high computational cost.
- The extreme performance gains OpenMC has achieved on GPUs is finally bringing within reach a much larger class of problems that historically were deemed too expensive to simulate using Monte Carlo methods.



CRK-HACC (courtesy Adrian Pope, Steve Rangel, Nick Frontiere)

ESP/HACC PI: **Katrin Heitmann**
 ECP/ExaSky PI: **Salman Habib**

- CRK-HACC simulates the formation of large-scale structures in the Universe over cosmological time.
- CRK-HACC employs n-body methods for gravity and a novel formulation of Smoothed Particle Hydrodynamics.
- CRK-HACC is a mixed-precision C++ code, with FLOPS-intense sections implemented using architecture-specific programming models in FP32 precision.

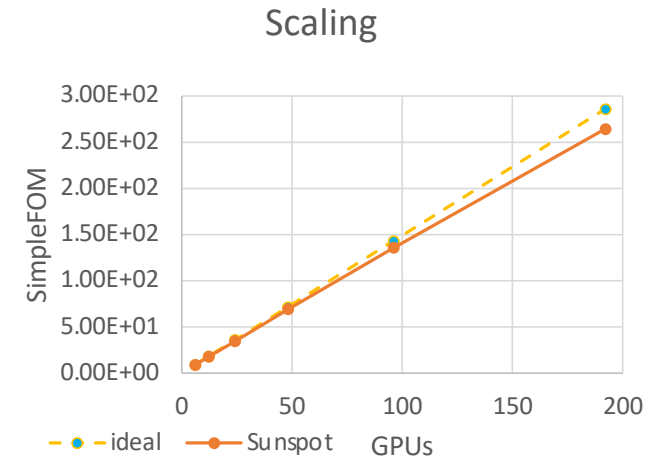
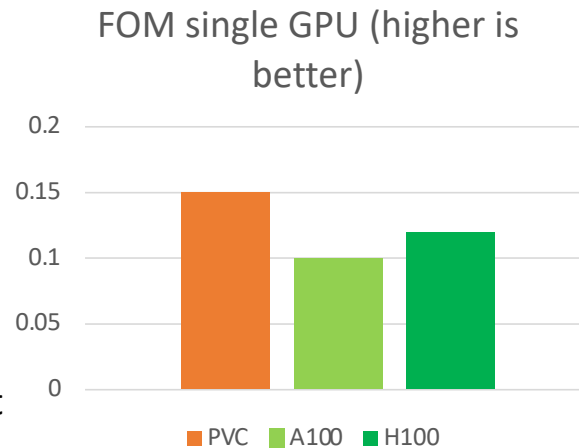
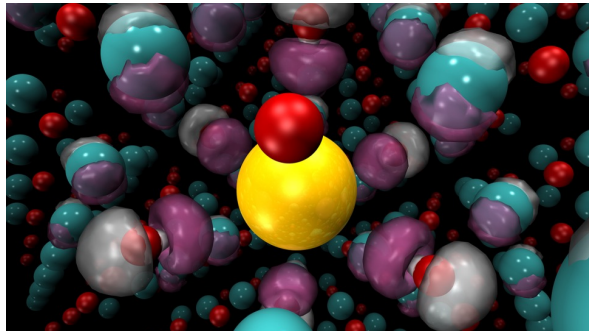


- CUDA and HIP are maintained as a single source with macros.
- SYCL kernels were translated from CUDA using SYCLomatic and custom LLVM-based tools, including optimizations for Intel GPUs.
- Figure-of-Merit (FOM) has units of particle-steps per second.
- Single GPU FOM problem used 33 million particles per GPU, and Intel PVC results are shown for both small (128) and large (256) General-purpose Register File (GRF) modes.
- Weak-scaling results are shown with the full application FOM, where the GPU represents roughly 80% of the total wall clock.

QMCPACK (courtesy Thomas Applencourt, Ye Luo, Jeongnim Kim)

ECP Project PI: Paul Kent

- QMCPACK, is a high-performance open-source Quantum Monte Carlo (QMC) simulation code.
- Science case: computing the quantum mechanical properties of materials with benchmark accuracy, including for energy storage and quantum materials.
- QMCPACK uses C++ and OpenMP target offload, plus wrappers (eg SYCL) around vendor optimized linear algebra.

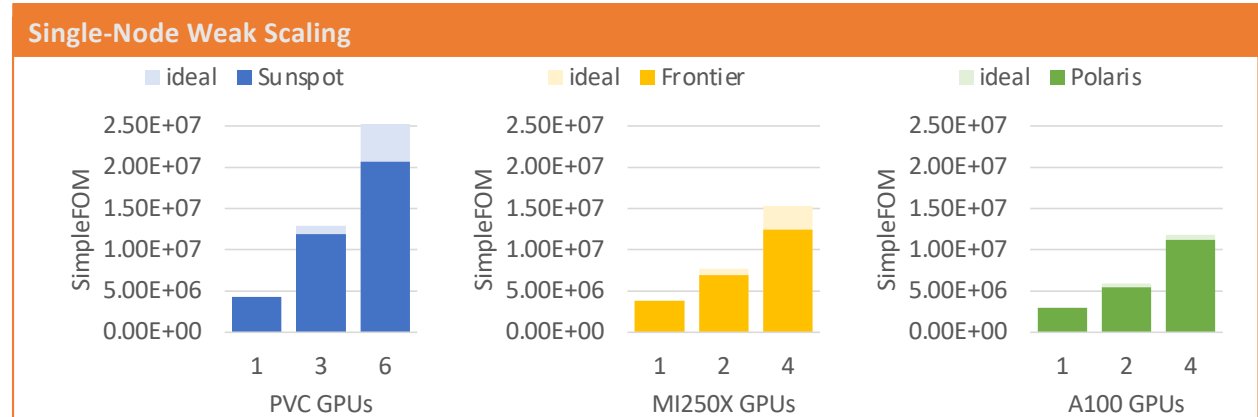
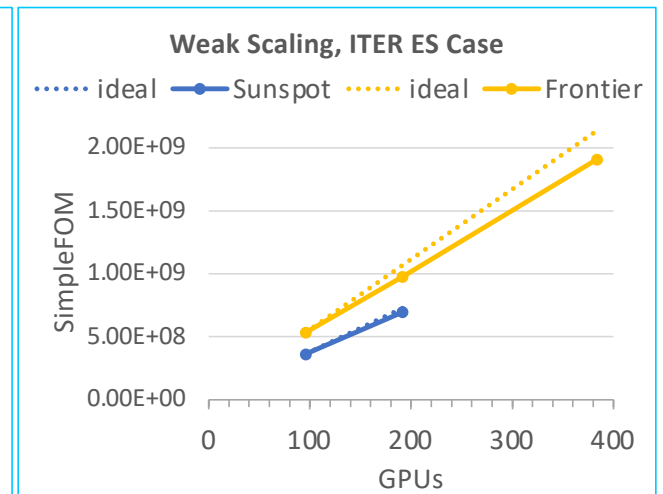
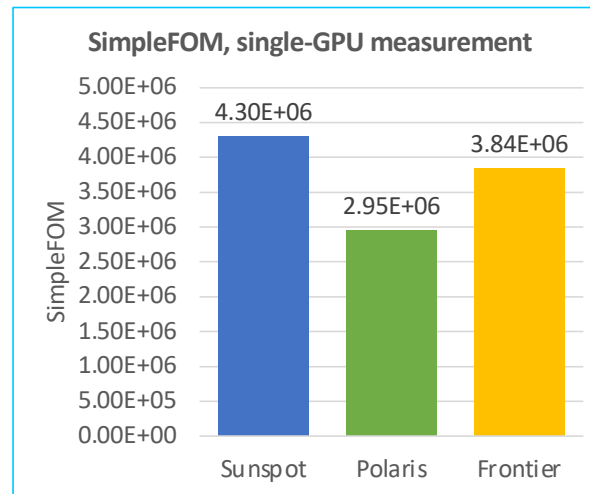
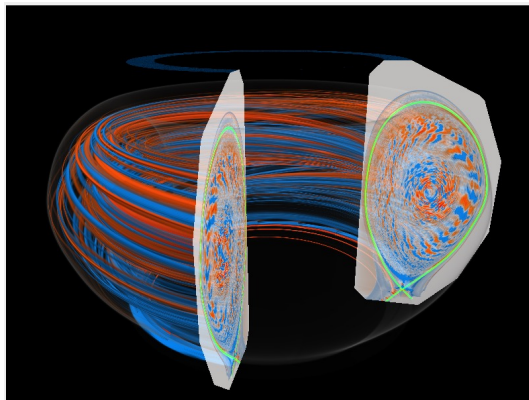


- Running `dmc-a512-e6144-DU64` problem. This simulates a supercell of nickel oxide with 6144 electrons and 512 NiO atoms total.
- Intel® Data Center GPU Max Series: 2 MPI ranks per GPU, 8 Walkers per rank, 64 GB of HBM per stack. Using Intel(R) oneAPI DPC++/C++ Compiler 2022.12.30
- A100 (40GB): 1 MPI Rank, 7 Walkers. LLVM15 compiler.
- H100: llvm/clang 17, cuda 11.8): 1 MPI Rank, 7 Walkers
- The Figure Of Merit (FOM) measure is throughput (walker moves/second). Higher is better.

XGC (courtesy Tim Williams, Aaron Scheinberg)

ESP Project PI: CS Chang
ECP Project PI: Amitava Bhattacharjee

- Science case: Predict ITER fusion reactor plasma behavior with Tungsten impurity ions sputtered from the divertor
- Gyrokinetic particle-in-cell simulation of tokamak plasma using C++ and:
 - Kokkos/SYCL on Intel GPUs
 - Kokkos/HIP on AMD GPUs
 - Kokkos/CUDA on NVIDIA GPUs

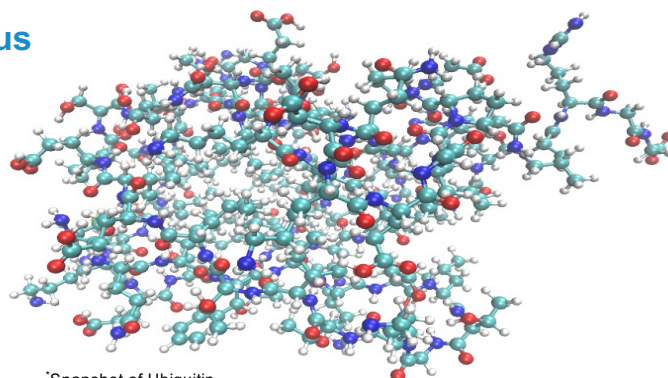


NWChemEx (Courtesy of Ajay Panyala)

<https://github.com/NWChemEx-Project>

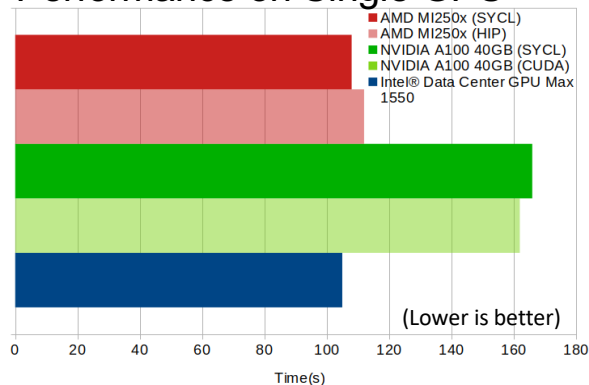
ESP & Project Project PI: Theresa Windus

- NWChemEx is a general purpose electronic structure code, which includes
 - Array of high-fidelity coupled cluster methods
 - Hartree-Fock, DFT, MP2 methods
 - Reduced-scaling DLPNO formulation
 - Molecular dynamics
- Programming models: C++, CUDA, HIP, SYCL
 - Communication frameworks: Global Arrays, UPC++, MADNESS
 - Tensor Contraction Engines: TAMM, TiledArray
- Key physics modules
 - DLPNO-CCSD(T)
 - Reduced-scaling implementation for GPU platforms



*Snapshot of Ubiquitin Protein

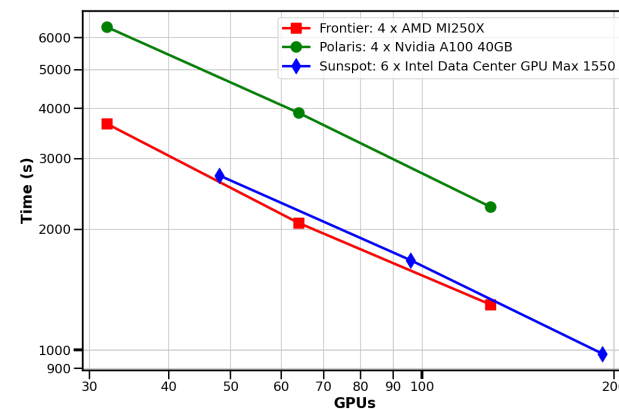
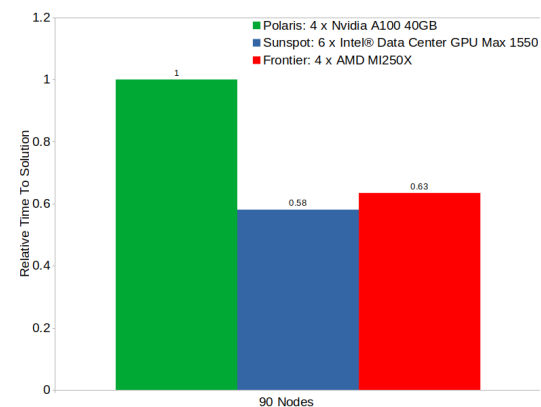
Performance on Single GPU



- Single GPU, Time in seconds for DLPNO-CCSD per iteration
- Performance of SYCL on NVIDIA & AMD were comparable with native CUDA & HIP respectively



Strong Scaling Performance on 90-nodes

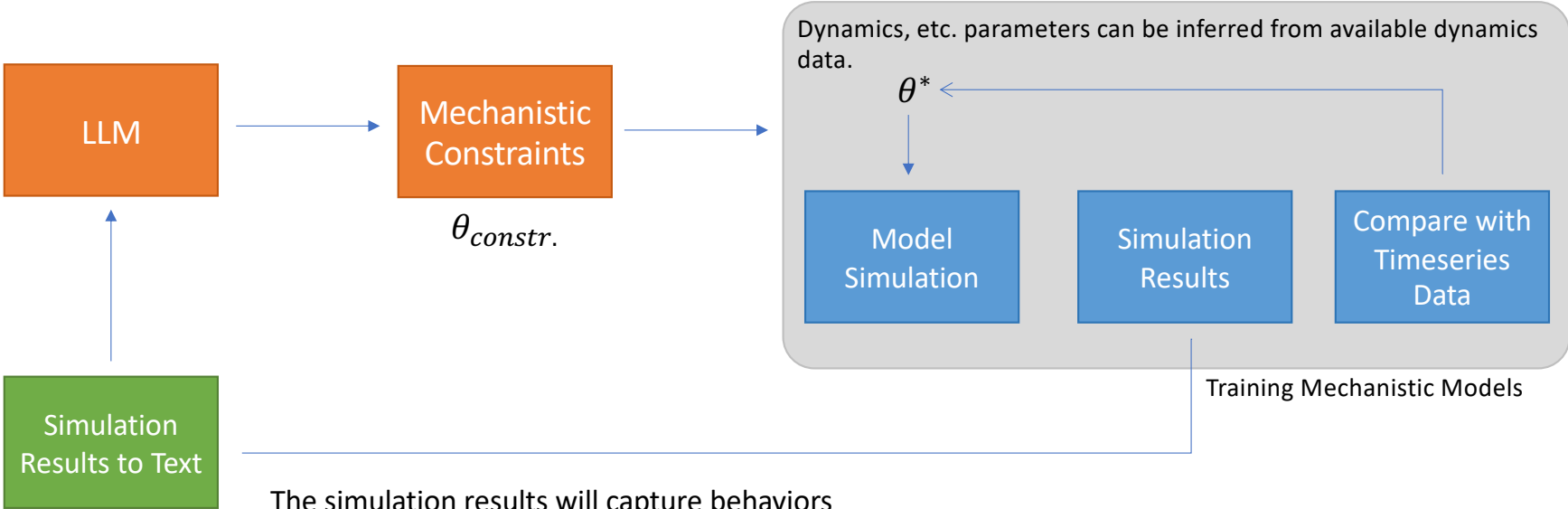


- Acknowledgment: Work performed by the NWChemEx team members without any architecture specific optimizations

AI Managed Sims

Language model outputs predictions and insights that guide the formulation of constraints for the mechanistic (or simulation) model.

The constraints can encode specific experimental conditions, predicted/hypothetical interactions, associations, etc.



The simulation results will capture behaviors (e.g., dynamics) that are not accessible by direct observation, but are supported by indirect observations, experiments, etc.

By translating these simulation results to text we provide new simulation-supervised dataset that would improve the LLM.

Prototyping this in the context of Radiation biology.. Cancer and space travel Are the motivations

The image features a large, bold, white capital letter 'A' centered in the middle. The background is a complex, glowing digital circuit board pattern. The lines of the circuitry are primarily in shades of blue and purple, with some yellow and orange highlights, giving it a futuristic, high-tech appearance. The overall aesthetic is clean and modern, typical of digital or artificial intelligence branding.

A

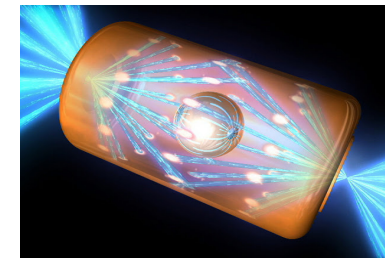
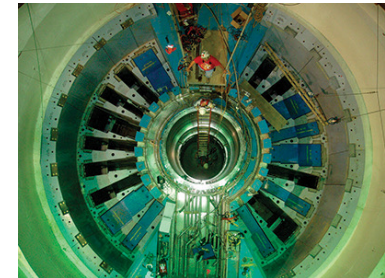


On the road to AI for Science

DOE's Unique Position for AI Leadership

- Operates the most capable computing systems and the world's largest collection of advanced experimental facilities
- Responsible for US nuclear security through deep partnerships across government
- Largest producer of classified and unclassified scientific data in the world
- Strongest foundation combining physical, biological, environmental, energy, mathematical and computing sciences
- Largest scientific workforce in the world
- Strong ties with private sector technology and energy organizations and stakeholders

Leadership in experimental facilities and supercomputers



AI for Science, Energy and Security

2019



What changed in three years?

- Language Models (e.g. ChatGPT) released
- Artificial image generation took off
- AI folded a billion proteins
- AI hints at advancing mathematics
- AI automation of computer programming
- Explosion of new AI hardware
- AI accelerates HPC simulations
- Exascale machines start to arrive

2022



2020 DOE Office of Science ASCR Advisory Committee report recommending major DOE AI4S program

Report posted here:

<https://www.anl.gov/ai-for-science-report>



U.S. DEPARTMENT OF
ENERGY

Office of
Science



Workshops organized on six crosscutting themes

AI for advanced properties inference and inverse design

Energy Storage
Proteins, Polymers,
Stockpile modernization

AI and robotics for autonomous discovery

Materials, Chemistry, Biology
Light-Sources, Neutrons

AI-based surrogates for high-performance computing

Climate Ensembles
Exascale apps with surrogates
1000x faster => Zettascale now

AI for software engineering and programming

Code Translation, Optimization
Quantum Compilation, QAlgs

AI for prediction and control of complex engineered systems

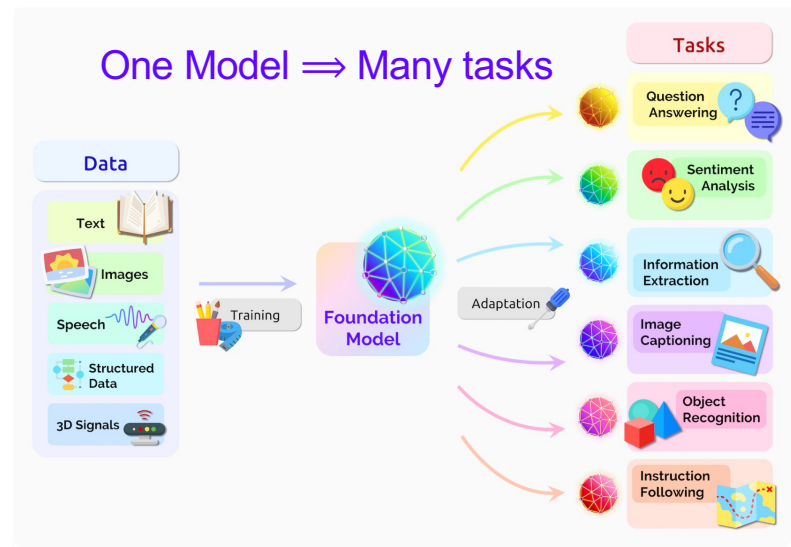
Accelerators, Buildings, Cities
Reactors, Power Grid, Networks

Foundation, Assured AI for scientific knowledge

Hypothesis Formation, Math
Theory and Modeling Synthesis,

Foundation Models — What are they?

- **Large scale model trained on large datasets from many sources** (text, papers, datasets, code, molecules, etc.)
- **Additional training to improve the human interaction experience** (e.g., ChatGPT-4)
- **Large models are remarkably flexible and exhibit emergent behaviors** (capable of tasks not originally trained to do)
- **Many hundreds of applications built on top**
- There are early efforts underway in DOE labs to create Foundation Models explicitly targeting scientific discovery



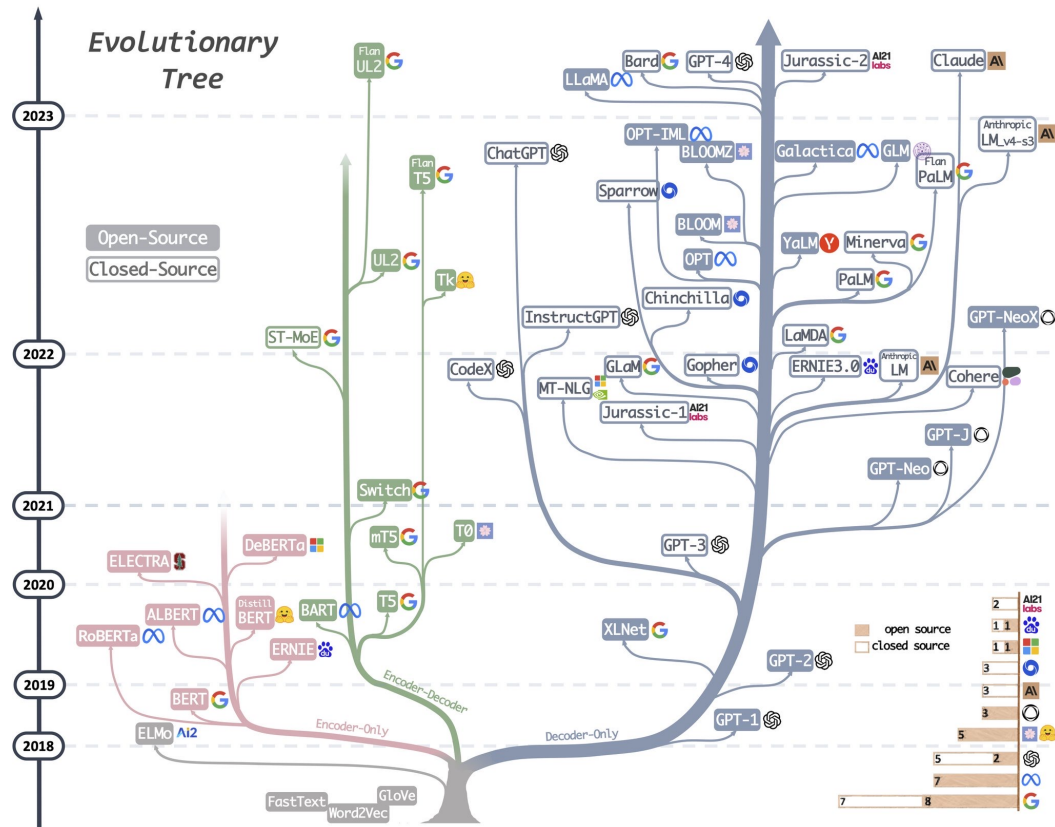
Trained on trillions of input "tokens" for many weeks on a large-scale computers

SOTA models (GPT-4) have about 1 trillion parameters (1% brain scale)



Since 2019 LLM model development has accelerated

Rapid Development of Large Language Models



Explosion of development of LLM based AI systems since 2019

Foundation Models are replacing narrow AI systems at a rapid pace

Foundation Models are the closest things that have yet been created that hint at the possibility of Artificial General Intelligence

Foundation Models for Science – Opportunities

- **FMs can summarize and distill knowledge** – extract information from million of papers into compact computing representation – **PPI networks, materials compositions, code kernels, biological function, etc.**
- **FMs can synthesize** – combine information from multiple sources – generate small programs for specific tasks – **quantum computing programs using QISkit & Cirq, derivations for applied physics, code for visualization and animation, etc.**
- **FMs can generate plans, solve logic problems** and write experimental protocols for robots – **powering self-driving labs, generate strategies for problem solving, and planning for testing hypotheses**
- **FMs with additional research, may be able to generate hypotheses to be tested and new theories for exploration** – a full-time shared scientific assistant that learns from across all of science is possible

Can ChatGPT be used to generate scientific hypotheses?
Yang Jeong Park^{1,2}, Daniel Kaplan¹, Zhichu Ren¹, Chia-Wei Hsu¹, Changhao Li¹, Haowei Xu¹, Sipei Li¹ and Ju Li^{1,4*}

¹ Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
² Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
³ Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot 7610001, Israel
⁴ Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
*Corresponding Author: liju@mit.edu

Abstract
We investigate whether large language models can perform the creative hypothesis generation that human researchers regularly do. While the error rate is high, generative AI seems to be able to effectively structure vast amounts of scientific knowledge and provide interesting and testable hypotheses. The future scientific enterprise may include synergistic efforts with a swarm of "hypothesis machines", challenged by automated experimentation and adversarial peer reviews.

In a university or research institute, a significant portion of fresh ideas arises out of discussions. Can talking to ChatGPT-4,¹ OpenAI's latest chatbot, create genuinely interesting scientific hypotheses?

In the past, only humans generated interesting hypotheses. Computers have been used to perform numerical simulations or even to prove theorems, like the four-color theorem in 1976². But making interesting laboratory-testable hypotheses with artificial intelligence (AI) seems far-fetched, until recently.

We are a collaborative group of experimental and theoretical researchers in physical sciences and engineering. Generative Pre-trained Transformer (GPT-4), released on March 14, 2023, is a large language model (LLM) significantly bigger than its predecessor GPT-3 released in 2020 (already with 1.75×10^{11} parameters). GPT-4 neural network was trained on a text corpus of books, webpages, academic papers from various disciplines, discussion forums, etc., up to September 2021. After experimenting with GPT-4 in our own research domains in materials chemistry, physics and quantum information, we find that ChatGPT-4 is knowledgeable, frequently wrong, and interesting to talk to. In other words, not unlike a college professor or a colleague.

To make everything concrete, our operative definition of "genuinely interesting scientific hypotheses" is (a) whether after a conversation, some experienced practitioner of a field can feel

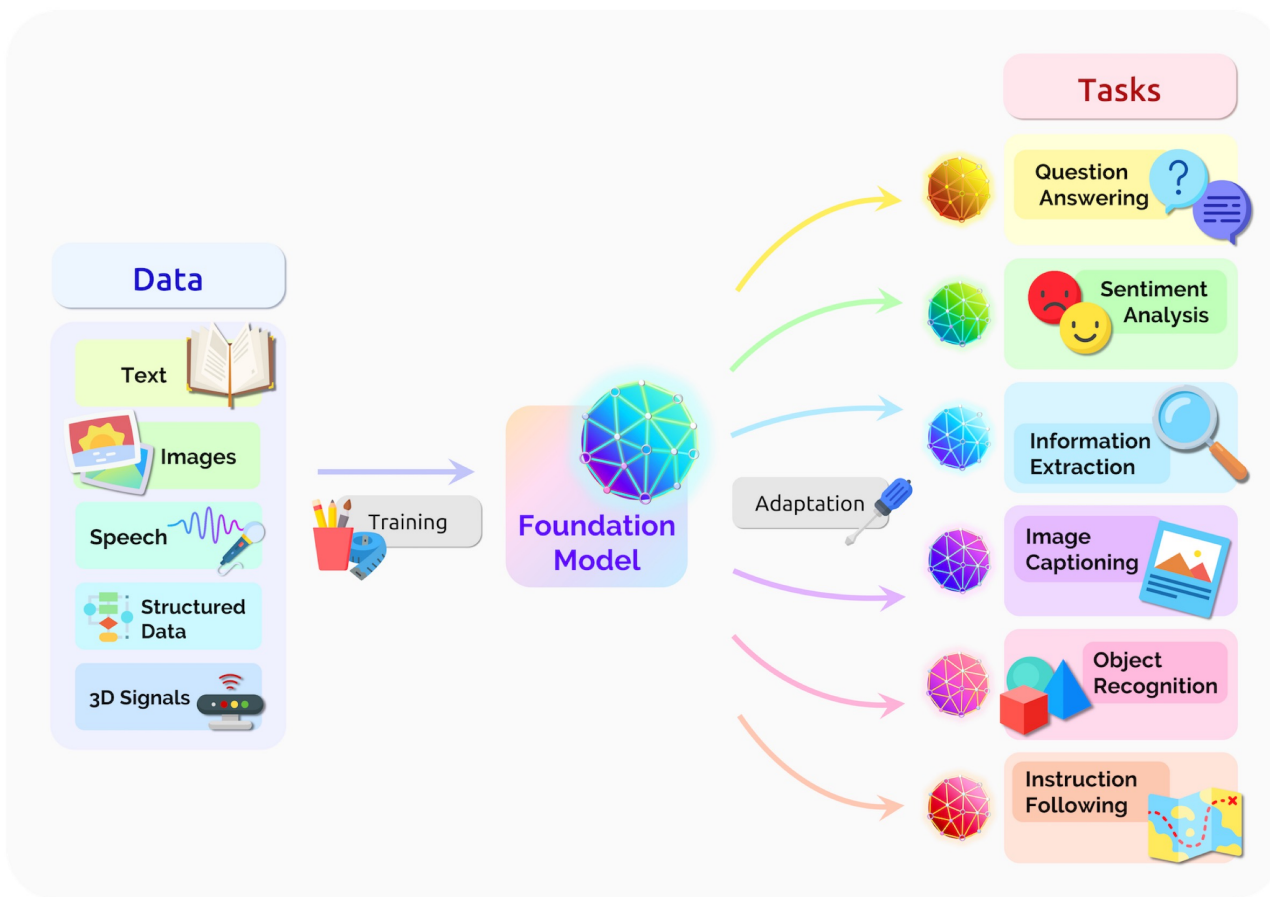
1

After experimenting with GPT-4 in our own research domains in materials chemistry, physics and quantum information, we find that ChatGPT-4 is knowledgeable, frequently wrong, and interesting to talk to. In other words, not unlike a college professor or a colleague. <https://arxiv.org/pdf/2304.12208.pdf>

Leveraging Community Efforts

Scientific & Engineering Datasets

Mathematics
Biology
Materials
Chemistry
Particle Physics
Nuclear Physics
Computer Science
Climate
Medicine
Cosmology
Fusion Energy
Accelerators
Reactors
Energy Systems
Manufacturing



Exemplar DOE Mission Tasks

Scientific Discovery

Digital Twins

Inverse Design

Code Optimization

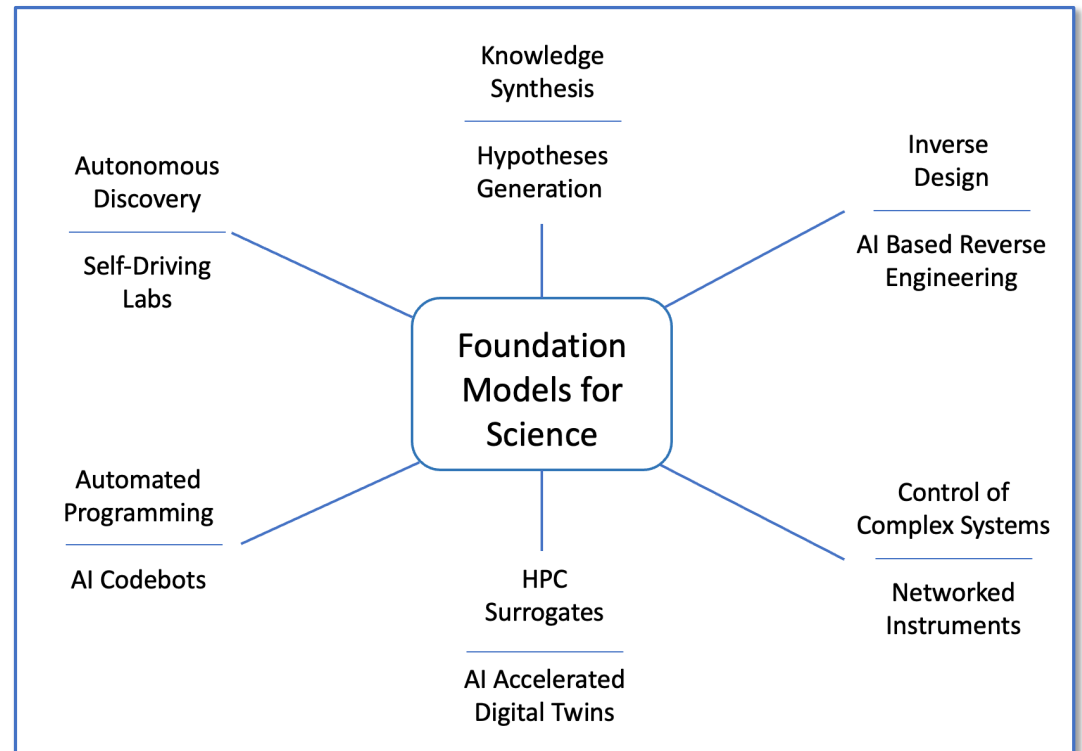
Accelerated Simulations

Autonomous Experiments

Secure Data Infrastructure

Co-Design

It is likely that many of the use cases we imagine in the AI4SES report can be driven directly or indirectly from sufficiently powerful Foundation Models

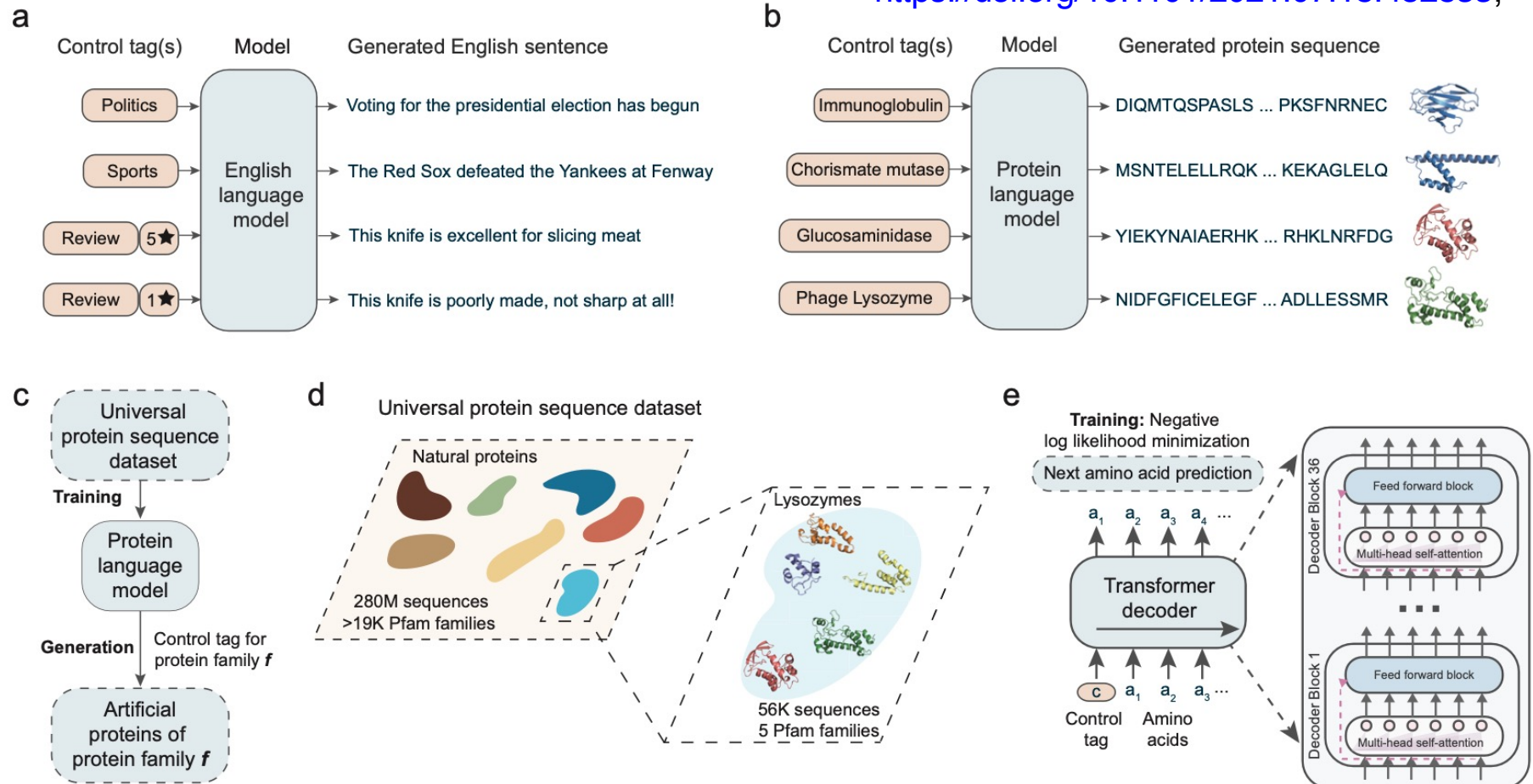


The background features a complex, glowing circuit board pattern in shades of blue, purple, and teal. The lines of the circuit are interconnected and radiate from a central point. A dark blue square is positioned in the center of the image, serving as a backdrop for the text.

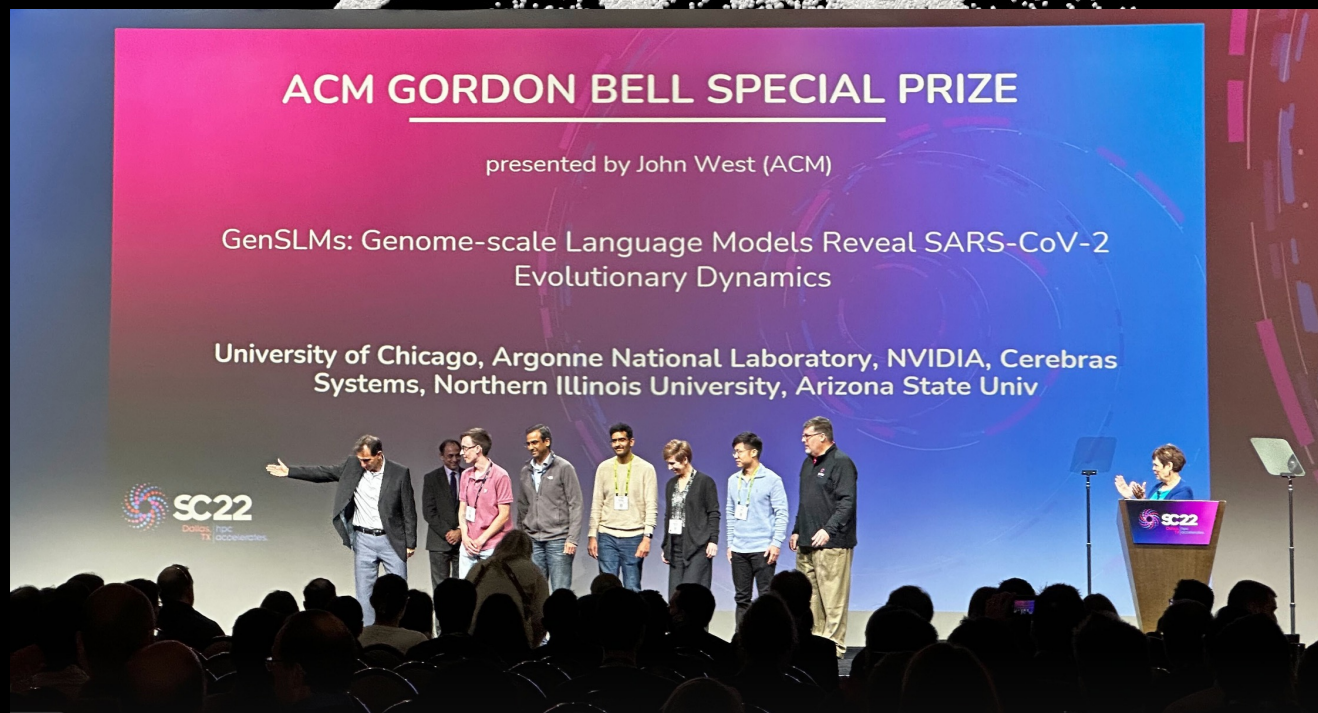
LLMs are already being used in many domains

Language models can Design Proteins

<https://doi.org/10.1101/2021.07.18.452833>;



GenSLM Foundation models reveal new biological insights on gene-level organization








The most capable models today are in the private sector (GPT-4, Claude, ChatGPT-3.5)

Large models with interesting emergent behavior

Assistant models in the wild

Assistant Models have been further trained to act as helpful chatbots

Rank	Model	Elo Rating	Description	License
1	 GPT-4	1274	ChatGPT-4 by OpenAI	Proprietary
2	 Claude-v1	1224	Claude by Anthropic	Proprietary
3	 GPT-3.5-turbo	1155	ChatGPT-3.5 by OpenAI	Proprietary
4	Vicuna-13B	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	Weights available; Non-commercial
5	Koala-13B	1022	a dialogue model for academic research by BAIR	Weights available; Non-commercial
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance	Apache 2.0
7	Oasst-Pythia-12B	928	an Open Assistant for everyone by LAION	Apache 2.0
8	ChatGLM-6B	918	an open bilingual dialogue language model by Tsinghua University	Weights available; Non-commercial
9	StableLM-Tuned-Alpha-7B	906	Stability AI language models	CC-BY-NC-SA-4.0
10	Alpaca-13B	904	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford	Weights available; Non-commercial
11	FastChat-T5-3B	902	a chat assistant fine-tuned from FLAN-T5 by LMSYS	Apache 2.0
12	Dolly-V2-12B	863	an instruction-tuned open large language model by Databricks	MIT
13	LLaMA-13B	826	open and efficient foundation language models by Meta	Weights available; Non-commercial

AI Accelerated Post-Exascale Ecosystem

Scientific & Engineering Datasets

Mathematics
Biology
Materials
Chemistry
Particle Physics
Nuclear Physics
Computer Science
Climate
Medicine
Cosmology
Fusion Energy
Accelerators
Reactors
Energy Systems
Manufacturing

Text and Code Corpora

General Text
Social Media
News
Humanities
History
Law
Digital Libraries
OSTI Archive
Scientific Journals
arXiv
Code repositories
Laboratory Notes
PubMed
Agency Archives

DOE and NNSA Exascale Systems
Common AI Software Frameworks
Responsible AI Techniques



Training



Training



Open Science Foundation Models

National Security Foundation Models

Tuned and Adapted Downstream Models

Exemplar DOE Mission Tasks

Scientific Discovery

Digital Twins

Inverse Design

Code Optimization

Accelerated Simulations

Autonomous Experiments

Secure Data Infrastructure

Co-Design

Integrated Research Infrastructure
Online Experimental Facilities
Strategic Partnerships



DOE is developing a concept for a large-scale program to implement the AI4SES vision

we call it

FASST: Frontiers of Ai for Science, Security and Technology

Integrated Program to Advance Trustworthy AI: Public-Private Partnerships that Include Labs, Academia and Industry

**Integrated science R&D for AI alignment,
trust and responsibility**

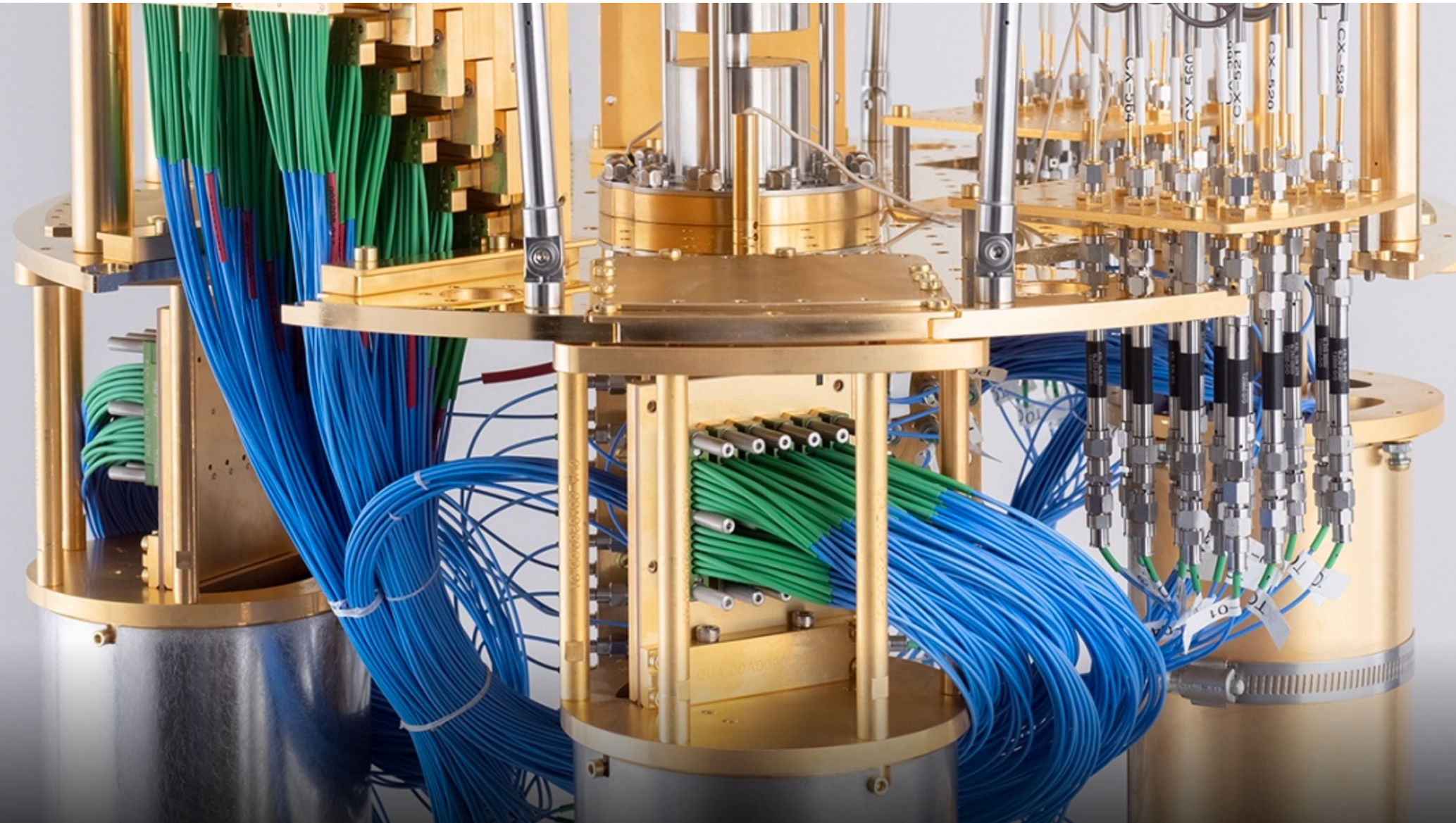
**Transformational hub-scale-centers on key AI4SES themes
strong ties to program grand challenges**

Crosscutting AI technologies

**Dedicated access to computing and
experimental facilities**

A large, white, serif letter 'Q' is centered on a dark background. The background is filled with a complex, glowing pattern of circuitry and data lines in shades of blue, purple, and teal. The lines are interconnected and form a dense, maze-like structure. There are also small, glowing circular nodes scattered throughout the circuitry. The overall effect is that of a futuristic, high-tech digital environment.

Q



QUANTUM is it going to contribute?

- How to resolve if quantum can contribute meaningfully to solving REAL problems faster than CLASSICAL supercomputers
- Current state small QC machines, unreliable, “circuit” model for programming, lack of error correction, lack of a good number of killer apps (and superpolynomial speed up candidates), ad hoc integration strategies
- Target problems (Chemistry, Factoring) appear to require order a million qubits and billions of gate operations
- Today’s systems are order 100 qubits and 100 gate operations

Possible real targets

- **Quantum Simulation (Hamiltonians)**, solving problems like many body electrons for larger systems than we can do classically, but that do not require data or extended systems of mechanics (atomic reactions for small molecules and ground states, but probably not proteins and drug binding)
- **Quantum Approximate Optimization Algorithms**, approximately solving *combinatorial* optimization problems with various constraints on density, etc.
- **Quantum physics exploration**, using quantum computers to explore QFT and related physics problems
- **Algorithms research**, the real impact may be simply from pushing on algorithms as hard as possible and seeing some flow from QC back to classical methods

Quantum computing enhanced computational catalysis

Vera von Burg,¹ Guang Hao Low,² Thomas Häner,³ Damian S. Steiger,³
Markus Reiher,^{1,*} Martin Roetteler,² and Matthias Troyer^{2,†}

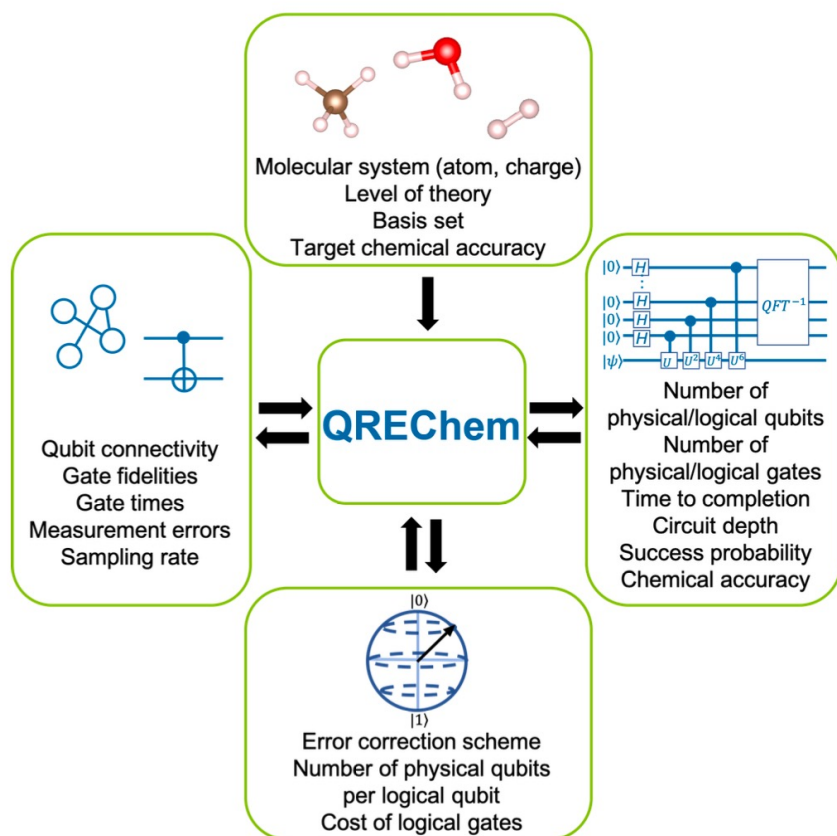
¹*Laboratorium für Physikalische Chemie, ETH Zürich,
Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland*

²*Microsoft Quantum, Redmond, Washington 98052, USA*

³*Microsoft Quantum, 8038 Zürich, Switzerland*

(Dated: March 5, 2021)

The quantum computation of electronic energies can break the curse of dimensionality that plagues many-particle quantum mechanics. It is for this reason that a universal quantum computer has the potential to fundamentally change computational chemistry and materials science, areas in which strong electron correlations present severe hurdles for traditional electronic structure methods. Here, we present a state-of-the-art analysis of accurate energy measurements on a quantum computer for computational catalysis, using improved quantum algorithms with more than an order of magnitude improvement over the best previous algorithms. As a prototypical example of local catalytic chemical reactivity we consider the case of a ruthenium catalyst that can bind, activate, and transform carbon dioxide to the high-value chemical methanol. We aim at accurate resource estimates for the quantum computing steps required for assessing the electronic energy of key intermediates and transition states of its catalytic cycle. In particular, we present new quantum algorithms for double-factorized representations of the four-index integrals that can significantly reduce the computational cost over previous algorithms, and we discuss the challenges of increasing active space sizes to accurately deal with dynamical correlations. We address the requirements for future quantum hardware in order to make a universal quantum computer a successful and reliable tool for quantum computing enhanced computational materials science and chemistry, and identify open questions for further research.



QREChem: Quantum Resource Estimation Software for Chemistry Applications

Matthew Otten^{1,†*}, Byeol Kang^{2,†}, Dmitry Fedorov³, Joo-Hyoung Lee², Anouar Benali³, Salman Habib³, Stephen Gray⁴ and Yuri Alexeev³

¹Materials and Microsystems Laboratory, HRL Laboratories, Malibu, CA 90265, USA

²School of Materials Science and Engineering, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

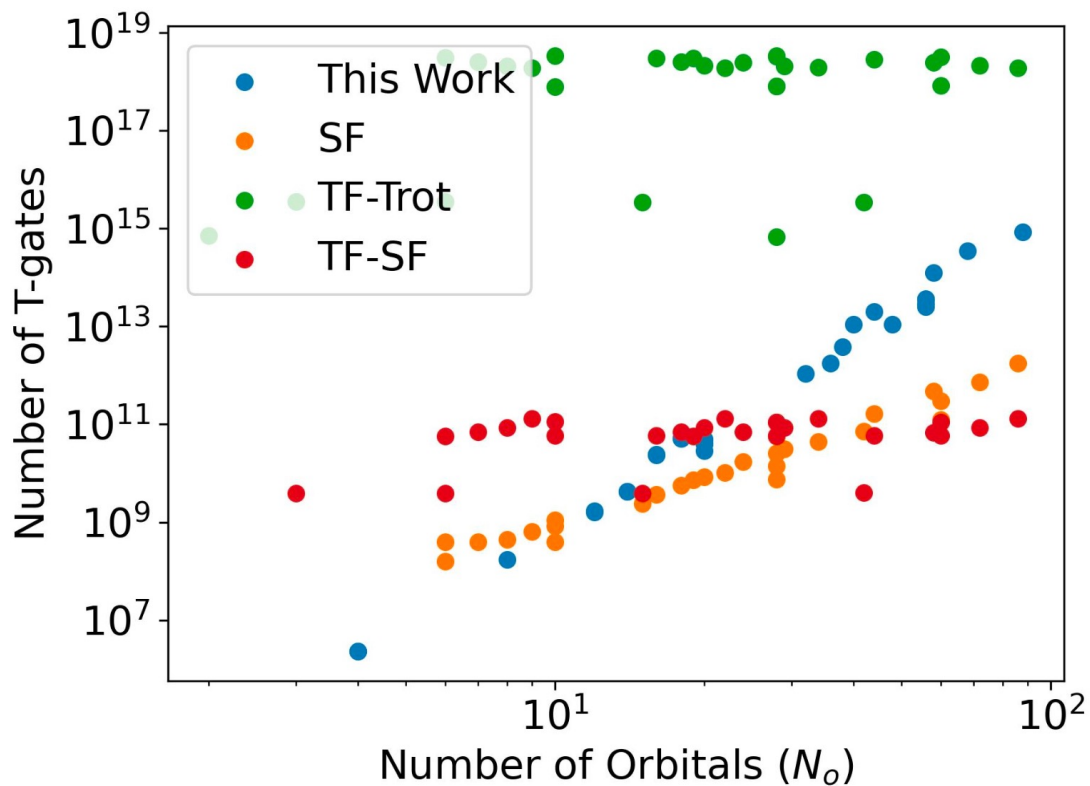
³Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA

⁴Center for Nanoscale Materials, Argonne National Laboratory, Lemont, IL 60439, USA

Correspondence*:
Corresponding Author
mjotten@hrl.com

2 ABSTRACT

3 As quantum hardware continues to improve, more and more application scientists have entered
4 the field of quantum computing. However, even with the rapid improvements in the last few
5 years, quantum devices, especially for quantum chemistry applications, still struggle to perform
6 calculations that classical computers could not calculate. In lieu of being able to perform specific
7 calculations, it is important have a systematic way of estimating the resources necessary to
8 tackle specific problems. Standard arguments about computational complexity provide hope that
9 quantum computers will be useful for problems in quantum chemistry but obscure the true impact
10 of many algorithmic overheads. These overheads will ultimately determine the precise point when
11 quantum computers will perform better than classical computers. We have developed QREChem
12 to provide logical resource estimates for ground state energy estimation in quantum chemistry



IBM's 100x100 Challenge

What is the 100×100 Challenge? In 2024, IBM plans to offer a tool capable of calculating unbiased observables of circuits with 100 qubits and depth-100 gate operations in a reasonable runtime.

Figure 2. Estimated total numbers of T gates for various algorithms over many molecules at many basis set levels. See text for the definitions of the algorithms.

Due to its many applications in chemistry and materials science, this problem is widely regarded as the “killer application” of future quantum computers, a view that was supported by our first rigorous resource estimate study for the accurate calculation of electronic energies of a challenging chemical problem.

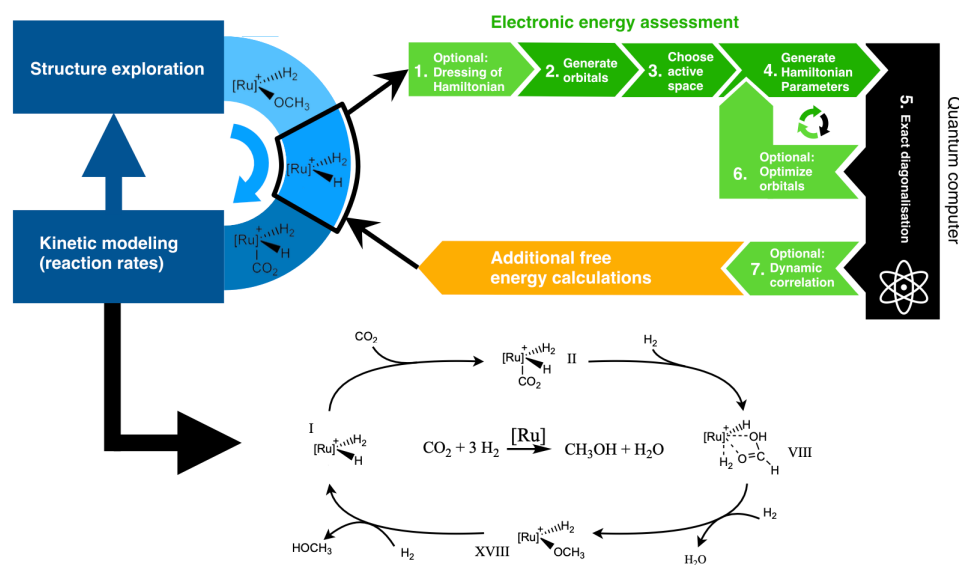


Figure 3. Protocol of computational catalysis with the key step of quantum computing embedded in black, which is usually accomplished with traditional methods such as CASSCF, DMRG, or FCIQMC (see text for further explanation).

Table I. Number of Toffoli gates for estimating an energy level to an error of 1 mHartree using a truncation threshold of $\epsilon_{in} = 1\text{mHartree}$ for the largest active spaces of structures in the catalytic cycle considered here. Our approach allows for a trade-off between the number of logical qubits required and the Toffoli count.

Structure	Orbitals	Electrons	R	M	α_{DF} /Hartree	Using fewer qubits Qubits	Using fewer Toffolis Toffolis/ 10^{10}	Using fewer Toffolis Qubits	Using fewer Toffolis Toffolis/ 10^{10}
I	52	48	613	23566	177.3	3400	1.3	6900	1.1
II	62	70	734	33629	374.4	4200	3.6	8400	3.1
II-III	65	74	783	38122	416.0	4400	4.5	8900	3.7
V	60	68	670	29319	371.1	4100	3.3	8200	2.9
VIII	65	76	794	39088	425.7	4400	4.6	8900	3.8
VIII-IX	59	72	666	29286	384.4	4000	3.4	8000	2.9
IX	62	68	638	28945	396.6	4200	3.5	8400	3.1
XVIII	56	64	705	29594	293.5	3700	2.5	7400	2.1

Table II. Comparison of our new double-factorization approach for H_{DF} applied to the FeMoco active site of nitrogenase ($N = 54$) with prior approaches based on Trotterization [11] or qubitization [22] using the unfactorized H or single-factorized H_{CD} Hamiltonian, and also for the VIII structure in the catalytic cycle ($N = 65$) where all examples apply the incoherent truncation scheme with the same threshold of $\epsilon_{in} = 1\text{mHartree}$.

Structure	Approach	α / Hartree	Terms	Qubits	Toffoli gates	Comments
FeMoco	Qubitization H_{DF}	300.5	1.3×10^6	3600	2.3×10^{10}	$\epsilon_{in} = 1\text{mHartree}$.
	Qubitization H_{DF}	296.9	2.8×10^5	3600	1.22×10^{10}	Optimistic $\epsilon_{in} = 73\text{mHartree}$.
	Trotterization H [11]	-	-	142	1.5×10^{14}	Optimistic Trotter number.
	Qubitization H [22]	9.9×10^3	4.4×10^5	5100	2.3×10^{11}	Truncation evaluated by CCSD.
VIII	Qubitization H_{CD} [22]	3.6×10^4	4.0×10^5	3000	1.2×10^{12}	Truncation evaluated by CCSD.
	Qubitization H_{DF}	425.7	2.5×10^6	4600	4.6×10^{10}	$\epsilon_{in} = 1\text{mHartree}$.
	Qubitization H	1.1×10^4	2.2×10^6	11000	9.3×10^{11}	$\epsilon_{in} = 1\text{mHartree}$.
	Qubitization H_{CD}	4.2×10^4	1.3×10^6	5800	2.1×10^{12}	$\epsilon_{in} = 1\text{mHartree}$.

Table III. Scaling of cost in our double-factorization approach with truncation threshold for the FeMoco active site of nitrogenase ($N = 54$). For comparison, the last line has $R = 200$ which matches that used Berry et al. [22].

ϵ_{in} / mHartree	Rank R	Eigenvectors M	Terms $M \times N$	α_{DF} / Hartree	Qubits	#Toffoli gates
1	567	2.4×10^4	1.30×10^6	300.5	3600	2.3×10^{10}
10	371	1.33×10^4	7.2×10^5	300.0	3600	1.67×10^{10}
100	178	4.2×10^3	2.3×10^5	295.8	3600	1.16×10^{10}
73	200	5.2×10^3	2.8×10^5	296.9	3600	1.22×10^{10}

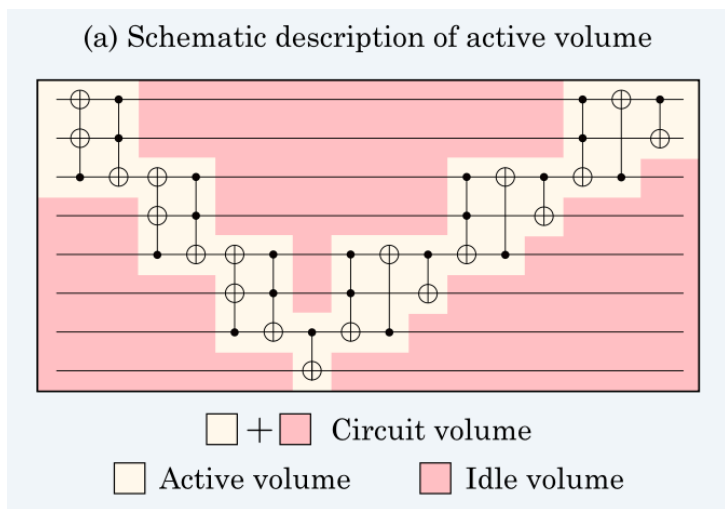
Thousands of qubits
Billions to trillions of gates

Shor's 2048 factoring 6 billion gates on 6200 logical qubits

Active volume: An architecture for efficient fault-tolerant quantum computers with limited non-local connections

Daniel Litinski and Naomi Nickerson
PsiQuantum, Palo Alto

Importantly, the architecture does not require all-to-all connectivity between N logical qubits. Instead, each logical qubit is connected to $O(\log N)$ other sites.



Assuming about 3K physical qubits per logical qubit in SC, Ion systems

arXiv:2211.15465v1 [quant-ph] 28 Nov 2022

Example algorithm: 2048-bit factoring algorithm with 500,000 lookup additions (6.1 billion T gates) on 6200 logical qubits

General-purpose architecture	
Old: Baseline architecture with 2D-local connectivity	New: Active-volume architecture with limited non-local connections
Cost function	
Circuit volume 3.8×10^{13}	Active volume 8.7×10^{11}
Superconducting qubit implementation with 1 μ s code cycle	
48 hours using 19 million physical qubits	54 minutes* using 19 million physical qubits
Trapped ion implementation with 1 ms code cycle	
5.4 years using 19 million physical qubits	37 days using 19 million physical qubits
Photonic implementation with 1 ns resource-state generation cycle	
48 hours using 9700 resource-state generators with 200 m fiber delays or 20 days using 970 resource-state generators with 2 km fiber delays or 200 days using 97 resource-state generators with 30 km free-space delays or 5.4 years using 10 resource-state generators with 300 km free-space delays	54 minutes* using 9700 resource-state generators with 200 m fiber delays or 8.9 hours using 970 resource-state generators with 2 km fiber delays or 3.7 days using 97 resource-state generators with 30 km free-space delays or 35 days using 10 resource-state generators with 300 km free-space delays



*if the reaction time is short enough

Figure 1: Resource estimates for the 2048-bit factoring algorithm described in Ref. [1] in a baseline architectures [2-6] and in the active-volume architecture described in this paper. More details are found in Appendix A.

QUANTUM is it going to contribute?

- We need machines with 1,000's of virtual-reliable qubits (1K-10K) able to run programs/circuits of depth $O(10^{10})$ - $O(10^{12}) \implies > 1M$ physical qubits and ~ 2 weeks of running at assuming $\sim \mu\text{s}$ - $\sim \text{ns}$ clocks
- We need algorithms for problems with better than quadratic speedups
- We need use cases where the value of the Quantum computation is greater than the cost of obtaining that result

Development Roadmap

Executed by IBM 
On target 

IBM Quantum

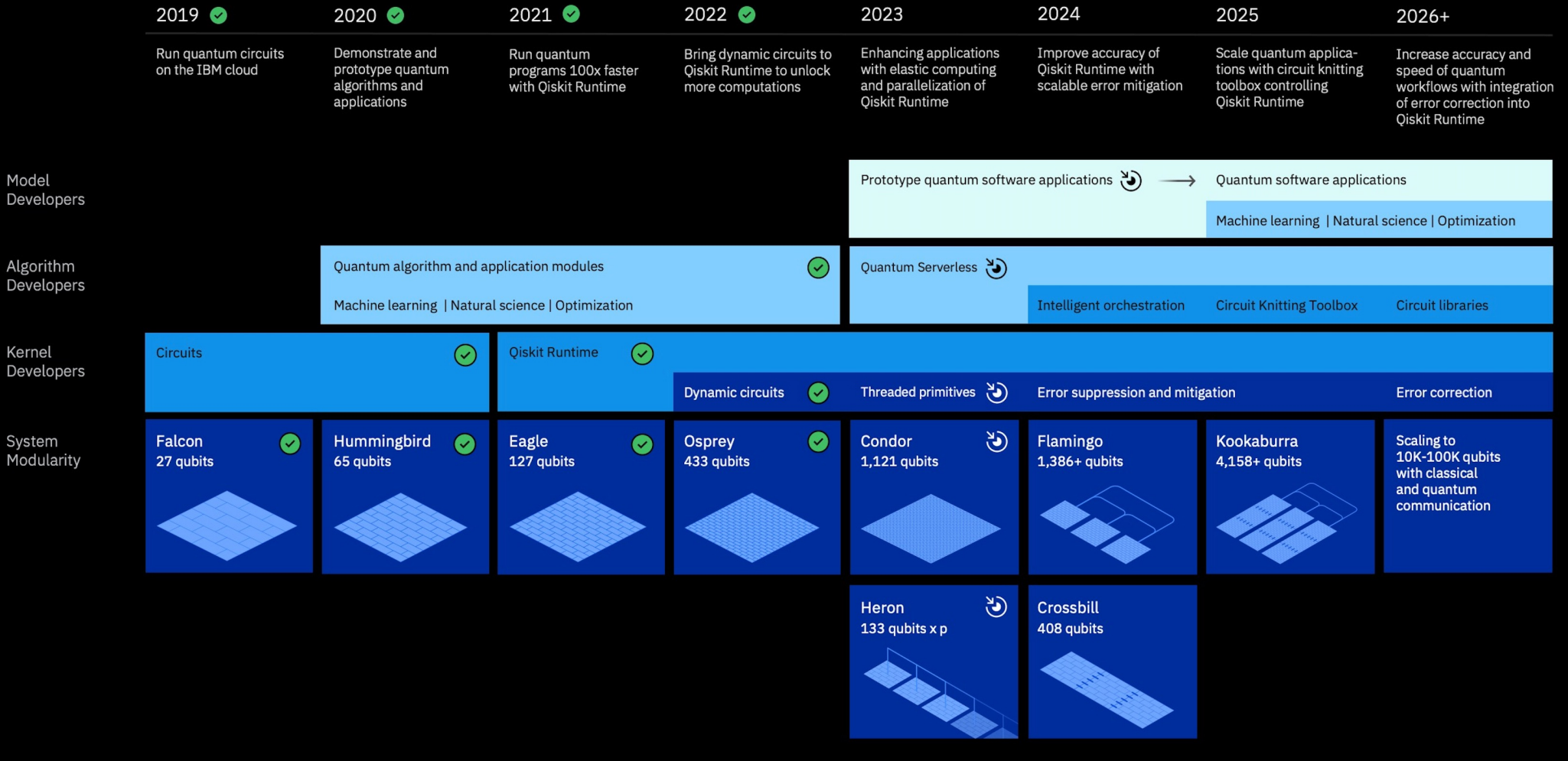
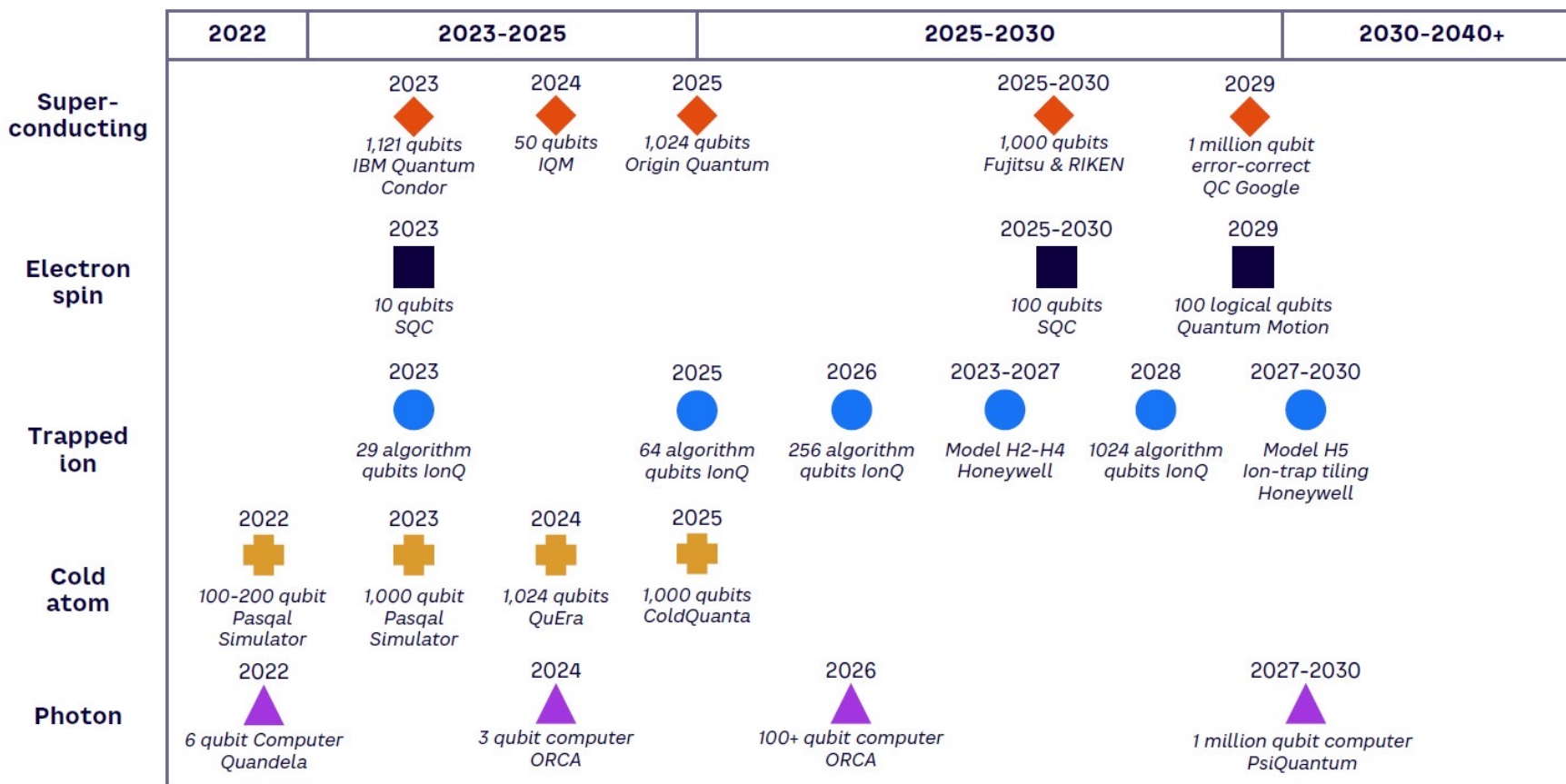


Figure 2. Quantum computing prototypes announced on vendor roadmaps



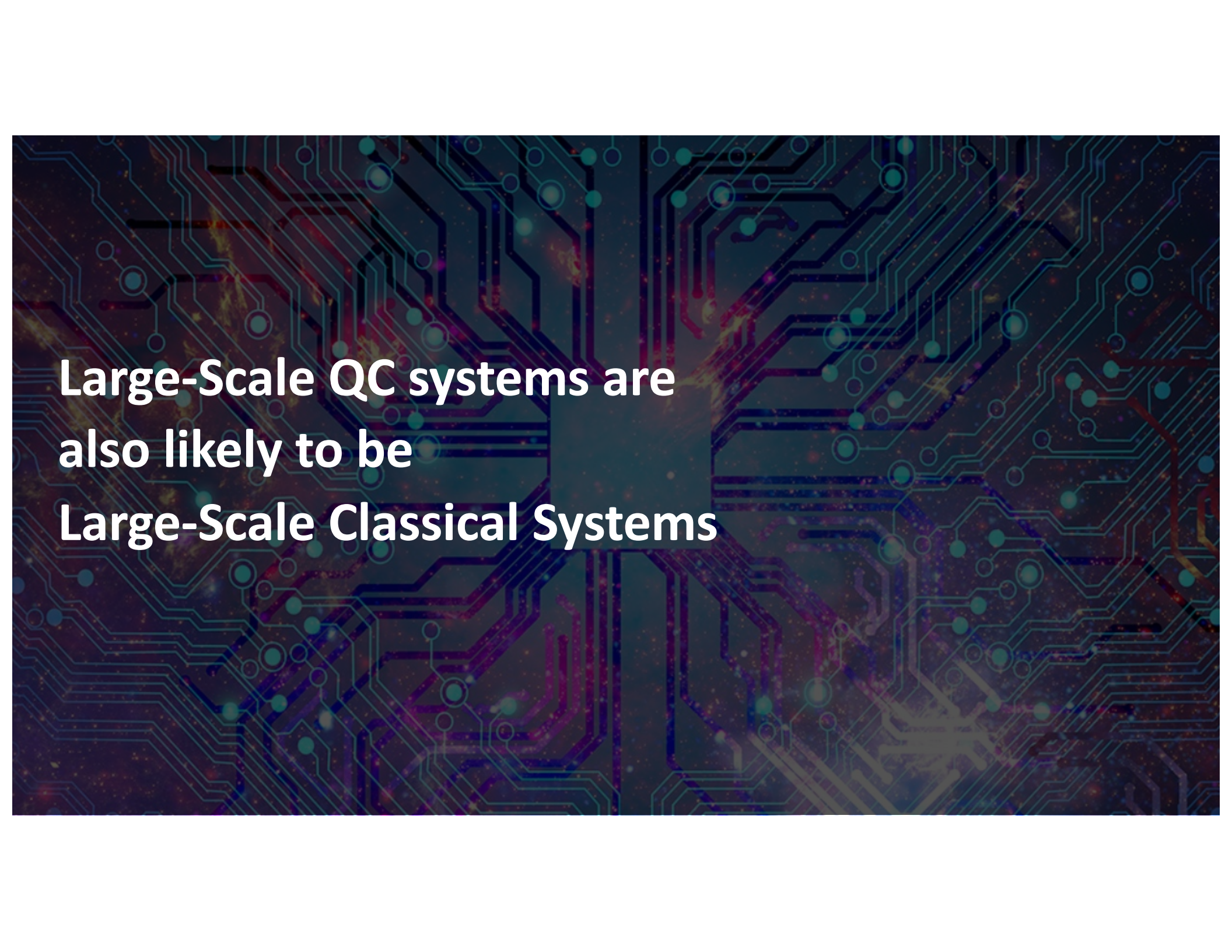
Source: Arthur D. Little, Olivier Ezratty

Can this be done?

1. **Massive search for fast (superpoly) Q algorithms** ($O(1000)$ computer scientists for a decade :-) + AI
2. **Government Commitments to field 2 or 3 > 1M qubit machines** each leveraging a different approach to qubits, scaling, reliability and fault tolerance (\$BILLION dollar machines)
3. **Scale** \Rightarrow Novel ways to connect and support active circuits, restrictive sets of operation, limitations on entanglement etc.
4. **Fab Paths** \Rightarrow Superconducting, photons, Ions/Atoms, dots, Majorana, etc.
5. **Building Blocks and Interfaces** \Rightarrow qubits, local/non-local comms, memory, control, classical interfaces etc.
6. **Leadership QC systems are likely to be embedded in leadership class classical machines** as all QC programs are hybrid

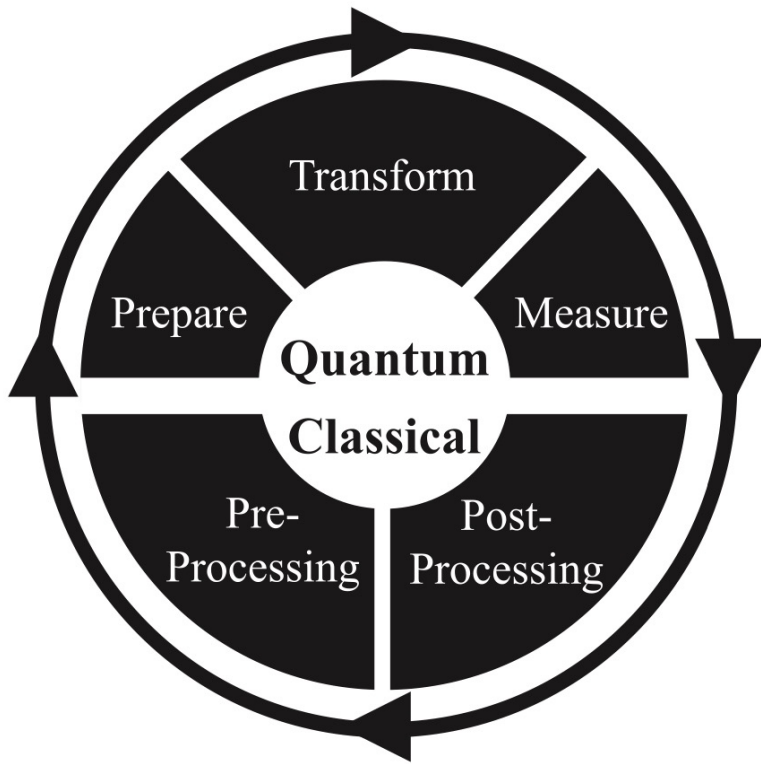
We need are more “exponential” algorithms

- Loosely speaking quantum algorithms fall into two broad classes, those that provides a superpolynomial or exponential speedup (e.g. Shor's) and those that are polynomial or less (e.g. Grovers)
- Without superpolynomial speedup it will be hard to beat classical machines for any real problem instance [see recent papers]
- Many algorithms also have the challenge that they assume data is already present in a “state prepared” and superimposed form
- Without breakthrough in quantum data loading, many applications that need to process real data at scale (AI, Operations Research, Biology, etc.) will be bottlenecked on loading data which is $O(n) + O(\dots)$ to just load data and will be limited by coherence times to load data



**Large-Scale QC systems are
also likely to be
Large-Scale Classical Systems**

Quantum Computing Paradigms -- Mixed



```

define a quantum operator
name the operator dft (for discrete Fourier transform)
the operator will act on a quantum register named q
operator dft (qreg q) {
classical iteration
  const n=#q;           number of qubits in q
  int i; int j;         classical variables for loop indices
  for i=0 to n-1 {      outer loop
    for j=0 to i-1 {    inner loop
      CPhase(2*pi/2^(i-j+1),  conditional phase rotation
              q[n-i-1] & q[n-j-1]);  angle of phase rotation
                                      rotate if state of these qubits is 11
    }
    Mix(q[n-i-1]);      place qubit in state
                        of maximum superposition
  }
  flip(q);              reverse order of qubits
}
  
```

A program for computing the discrete Fourier transform is written in the programming language QCL. The language combines elements that require a classical—that is, nonquantum—computer (*pink*) with operations that are unique to quantum processors (*blue*).

American Scientist, Volume 102

A large, white, stylized letter 'Z' is centered on a dark background. The background is filled with a complex, glowing pattern of circuit traces and nodes in shades of blue, purple, and teal, set against a dark, starry space-like texture. The traces form a dense, interconnected network, with some nodes highlighted in bright blue or purple. The overall aesthetic is futuristic and technological.

Z

ZETTASCALE – or how we continue to make progress

- Nominal goal would be a system 1000x (fp64) today's EXA machines
- But EXA is not one thing (fp64 for classic sim, fp32, tf32, bfp16 for AI training, fp8, int8, int4 for AI inference) and of course int4 >> fp64
- Current EXA design points are 20x -- 40x faster for AI inference compared to fp64 GEMM
- MANY MANY OPEN ISSUES (power, bandwidth, memory, interconnect)
 - representation flexibility – from 128-bit to 1-bit formats
 - effective scalar-vector-matrix-tensor modes
 - Throughput vs latency, memory hierarchy, levels of aggregation micro-macro
 - do ISAs matter? can we do better than x86 or ARM or is even necessary
 - Can we cheat on software using JAX like HLL binding to LL-ops



Some very high-level questions to consider

How much high-precision capability do we need for scientific problems?

Will AI driven surrogates be the dominate approach for simulations in the 2030's?

Can we improve on sustainability of systems?

We are not the only ones thinking about Zetta

Perspective:

Moving from exascale to zettascale computing: challenges and techniques*

Xiang-ke LIAO, Kai LU^{†‡}, Can-qun YANG, Jin-wen LI, Yuan YUAN, Ming-che LAI,
Li-bo HUANG, Ping-jing LU, Jian-bin FANG, Jing REN, Jie SHEN

College of Computer, National University of Defense Technology, Changsha 410073, China

[†]E-mail: kailu@nudt.edu.cn

Received Aug. 16, 2018; Revision accepted Sept. 14, 2018; Crosschecked

Abstract: High-performance computing (HPC) is essential for both traditional enabling scientific activities to make progress. With the development of high-performance that exascale computing will be put into practice around 2020. As Moore's law approaches, computing will face severe challenges when moving from exascale to zettascale, making a vital period to develop key HPC techniques. In this study, we discuss the challenges and techniques with respect to both hardware and software. We then present a perspective of future HPC revolution, leading to our main recommendations in support of zettascale computing.

Table 1 Zettascale metrics

Metric	Value
Peak performance	1 Zflops
Power consumption	100 MW
Power efficiency	10 Tflops/W
Peak performance per node	10 Pflops/node
Bandwidth between nodes	1.6 Tb/s
I/O bandwidth	10–100 PB/s
Storage capacity	1 ZB
Floor space	1000 m ²

Zettascale System Metrics

NUDT group proposes
Zettascale by 2035

Table 1 Zettascale metrics

Metric	Value
Peak performance	1 Zflops
Power consumption	100 MW
Power efficiency	10 Tflops/W
Peak performance per node	10 Pflops/node
Bandwidth between nodes	1.6 Tb/s
I/O bandwidth	10–100 PB/s
Storage capacity	1 ZB
Floor space	1000 m ²

PROGRESS REQUIRED

- 600x Frontier (58% CAGR)
- 3.4x Frontier (9% CAGR)
- 200x Frontier (46% CAGR)
- 66x Frontier (=> 10x nodes)
- 16x Frontier (22% CAGR)
- 1000x Frontier (64% CAGR)
- 1000x Frontier (64% CAGR)
- 2x Frontier (5% CAGR)

Moving from exascale to zettascale computing: challenges and techniques. Frontiers of Information Technology & Electronic Engineering, 19(10):1236-1244.
<https://doi.org/10.1631/FITEE.1800494> Front Inform Technol Electron Eng

Some 2028 and 2032 Planning Targets

2028 – 10 EF (sim fp64) and **>100 EF** (AI bfp16 or fp8)

2032 – 50 EF (sim fp64) and **>1000 EF** (AI bfp16 or fp8)

8x	Aurora	
	18.72	FP64
	18.72	FP32
	37.43	FP16
	149.72	FP32-m
	299.45	BFP16-m
	598.89	int8-m

A few questions we are pondering

- How achievable are these targets given roadmaps and vendor plans?
- Will AI accelerators (distinct from GPUs) make sense to integrate into future nodes or as sub-clusters?
- When will or if quantum computing accelerators intersect mainstream supercomputing? (IBM plans 100k qubits in 2033?)

Summary: Post-Exascale Directions

Push towards AI4S and hybrid AI/simulations

Data Driven Methods in General and AI for Science
ML Acceleration "AI surrogates" \Rightarrow 1000x \Rightarrow more
End-to-End AI alternatives to classical simulations

A long-term push on hardware towards Zettascale (2038 plausible)

1000x performance improvement over today's Exascale Systems

Probably in 3 or 4 steps of factors of 5x-10x over 15-20 years

(Exascale was $>$ 10 years from Petascale) Z is harder than E

Special purpose hardware for specific problems (AI, QC, etc.)

Brain scale AI ($>$ 100Trillion parameters) \Rightarrow Scientific AI

QC for Quantum Chemistry \Rightarrow Key Energy/Environment Challenges

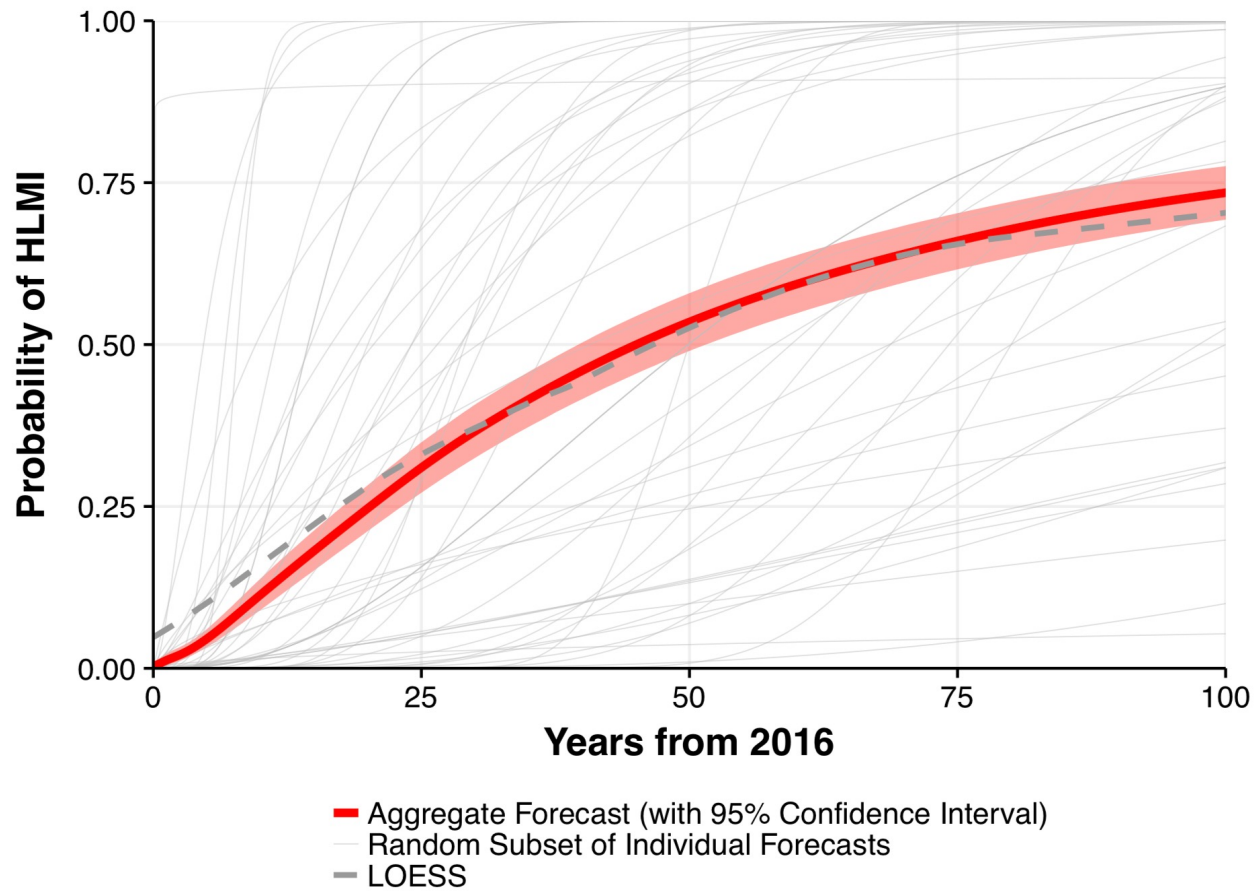


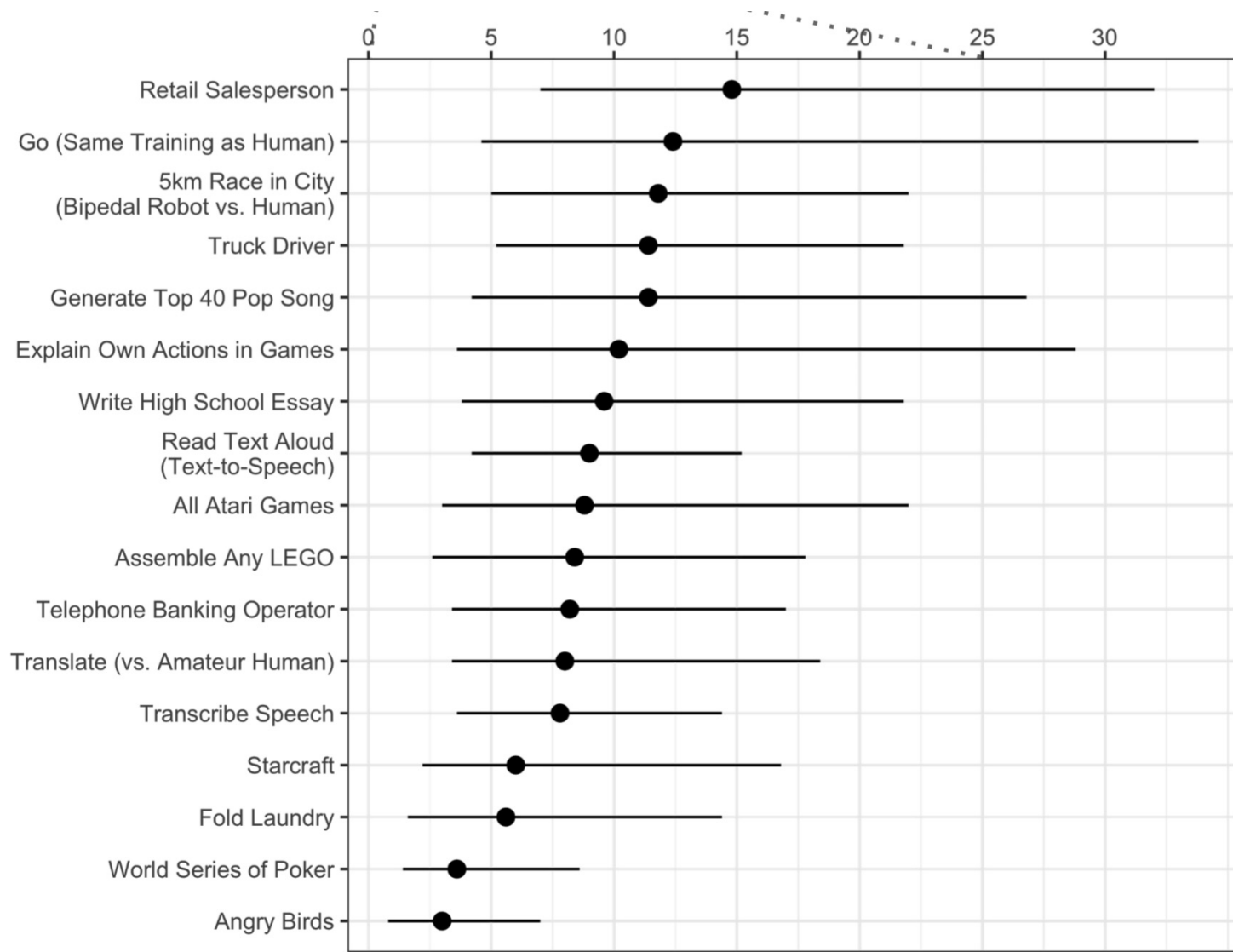
EAZQ

An aerial photograph of the Argonne National Laboratory campus, showing various buildings, parking lots, and green spaces. The image is overlaid with a semi-transparent dark layer. In the center, the word "Argonne" is written in a large, white, serif font. To its right is the Argonne logo, a white geometric shape consisting of three overlapping triangles forming a hexagon. Below "Argonne" and the logo, the words "NATIONAL LABORATORY" are written in a smaller, white, sans-serif font.

Argonne 
NATIONAL LABORATORY

Time for high-level machine intelligence





“Hard” jobs and AI

